Revista Brasileira de
Risco e Seguro

Modelagem de Seguro Desemprego: uma Abordagem Via Modelo Logístico para a Estimação do Risco de Perda da Renda de um Empregado Celetista

Gabriel Fonseca da Silva
Estatístico e bacharel em Ciências Atuariais pela Universidade do Estado do Rio de Janeiro (UERJ), pós-graduado em Gestão Estratégica pelo Instituto A Vez do Mestre/Universidade Candido Mendes e MBA em Gestão de Sistemas de Informação com o SAP pelo Instituto Infnet. Atua como estatístico na Eletrobras desde 2009, na área de gestão de pessoas.
gabrielfonsecarj@gmail.com

Thiago Barata Duarte
Graduado em Ciências Atuariais pela Universidade do Estado do Rio de Janeiro (2009), mestre em Engenharia Elétrica pelo Departamento de Engenharia Elétrica (DEE) na Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), com ênfase em Métodos de Apoio à Decisão (2015). Analista Técnico da Superintendência de Seguros Privados (Susep) onde atuou como Coordenador da Coordenação de Monitoramento de Riscos (CORIS) e atualmente exerce o cargo de Assessor da Diretoria de Normas e Habilitação das Operadoras (DIOPE) na Agência Nacional de Saúde Suplementar (ANS), tendo experiência na área de Probabilidade e Estatística nas Ciências Atuariais e Finanças, atuando principalmente nos seguintes campos: avaliação de riscos e modelos de mensuração de capital. É professor de Ciências Atuariais no Instituto de Matemática e Estatística (IME) da UERJ desde 2015, além de lecionar em outros cursos desde 2010.
tbarata85@gmail.com

Resumo

Este estudo teve como principal objetivo desenvolver um modelo estatístico para calcular a probabilidade de um empregado celetista ser demitido sem justa causa e definir o seu uso na tarifação de um seguro desemprego. Para isso foi aplicada uma regressão logística em uma base de dados de empregados celetistas do município do Rio de Janeiro para o ano de 2016, construída a partir dos microdados da Relação Anual de Informações Sociais (RAIS), disponibilizados em 2018 pelo extinto Ministério do Trabalho. A análise exploratória dos dados mostrou que o risco da perda do emprego tende a diminuir conforme o nível de escolaridade aumenta, não há grandes diferenças entre os níveis de demissões em relação ao sexo e à cor, e a construção civil foi o setor econômico com maior taxa de desligamentos sem justa causa dentre os demais. Por último, constatou-se que o modelo logístico gerado teve um ajuste razoável aos dados, com acurácia de 70% obtida pela curva ROC, e suas variáveis explicativas "sexo", "cor", "faixa etária", "grau de escolaridade", "portador de deficiência", "afastamento do trabalho", "tempo de emprego", "renda média em salários mínimos", "setor econômico" e "tamanho do estabelecimento em número de funcionários" foram significativas ao nível de 1%. A grande relevância deste trabalho reside no fato de ser utilizado um modelo simples e com bons resultados na tarifação desse produto com crescente demanda, em uma base de dados pública, ou seja, facilmente replicada por seguradoras sem massa de dados.

Palavras-chave

Seguro de pessoas. Perda da renda. RAIS. Modelos lineares generalizados. Regressão logística.

Sumário

Abstract

Unemployment Insurance Modeling: A Logistic Model Approach for Estimating the Risk of Income Loss for a Cashier Employee

Gabriel Fonseca da Silva
Statistician and bachelor in Actuarial Sciences from Rio de Janeiro State University (UERJ), postgraduate in Strategic Management from Instituto A Vez do Mestre / Candido Mendes University and MBA in Information Systems Management with SAP from Instituto Infnet. He has been a statistician at Eletrobras since 2009 in the area of people management.
gabrielfonsecarj@gmail.com

Thiago Barata Duarte
Graduated in Actuarial Sciences from the State University of Rio de Janeiro (2009), Master in Electrical Engineering from the Department of Electrical Engineering (DEE) at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), with an emphasis on Decision Support Methods (2015). Technical Analyst at the Superintendency of Private Insurance (Susep) where he served as Coordinator of the Coordination of Risk Monitoring (CORIS) and currently holds the position of Advisor to the Directorate of Standards and Qualification of Operators (DIOPE) at the National Supplementary Health Agency (ANS), having experience in the area of Probability and Statistics in the Actuarial Sciences and Finance, acting mainly in the following fields: risk assessment and capital measurement models. He has been a professor of Actuarial Sciences at the Institute of Mathematics and Statistics (IME) at UERJ since 2015, in addition to teaching in other courses since 2010.
tbarata85@gmail.com

Summary

The main objective of this study was developing a statistical model to calculate the probability of a formal employee unjustified dismissal and defyining its use in the pricing of an Unemployment Insurance. For this, a logistic regression was applied in a data base of employees of the city of Rio de Janeiro for the year 2016, which was constructed from the RAIS microdata, made available in 2018 by the extinct Ministry of Labour. The exploratory data analysis showed that the Unemployment Risk tends to decrease as the educational level increases, there are no relevant difference between the levels of layoffs in relation to gender and color, and construction was the economic sector with the highest rate of unjustified dismissal among others. Finally, it was found that the logistic model generated a reasonable data adjust, with 70% ROC curve accuracy, and its explanatory variables ‘gender’, ‘color’, ‘age group’, ‘educational level’, ‘disabled workers’, ‘work departures’, ‘time of employment’, ‘average income in minimum salaries’, ‘economic sector’ and ‘size of establishment in number of employees’ were significant at the level of 1%. The great relevance of this work arises from the fact that was used a simple model, with good results, in a public database, and therefore it could be easily replicated by insurers without enough data in the pricing of this product with increasing demand.

Keywords

People insurance. Loss of income. RAIS. Generalized Linear Models. Logistic Regression.

Contents

Sinopsis

Modelo de seguro de desempleo: un enfoque de modelo logístico para estimar el riesgo de pérdida de ingresos para un empleado

Gabriel Fonseca da Silva
Estadístico y licenciado en Ciencias Actuariales de la Universidad Estatal de Río de Janeiro (UERJ), postgrado en Gestión Estratégica del Instituto A Vez do Mestre / Universidad Candido Mendes y MBA en Gestión de Sistemas de Información con SAP del Instituto Infnet. Ha sido estadístico en Eletrobras desde 2009 en el área de gestión de personas.
gabrielfonsecarj@gmail.com

Thiago Barata Duarte
Graduado en Ciencias Actuariales por la Universidad Estatal de Río de Janeiro (2009), Máster en Ingeniería Eléctrica por el Departamento de Ingeniería Eléctrica (DEE) en la Pontificia Universidad Católica de Río de Janeiro (PUC-Río), con énfasis en Métodos de Apoyo a la Decisión (2015) Analista técnico en la Superintendencia de Seguros Privados (Susep) donde se desempeñó como Coordinador de la Coordinación de Monitoreo de Riesgos (CORIS) y actualmente ocupa el cargo de Asesor de la Dirección de Normas y Calificación de Operadores (DIOPE) en la Agencia Nacional de Salud Complementaria (ANS), con experiencia en el área de Probabilidad y Estadística en Ciencias Actuariales y Finanzas, actuando principalmente en los siguientes campos: evaluación de riesgos y modelos de medición de capital. Ha sido profesor de Ciencias Actuariales en el Instituto de Matemáticas y Estadística (IME) en UERJ desde 2015, además de enseñar en otros cursos desde 2010.
tbarata85@gmail.com

Resumen

Este estudio tuvo como principal objetivo desarrollar un modelo estadístico para calcular la probabilidad de que un empleado celetista fuera despedido sin justa causa. Para ello, se aplicó una regresión logística en una base de datos de empleados celetistas del Municipio de Río de Janeiro para el año 2016, que fue construida a partir de los microdatos de la RAIS, puestos a disposición en 2018 por el extinto Ministerio del Trabajo. El análisis exploratorio de los datos mostró que la pérdida del empleo tiende a disminuir conforme aumenta el nivel de escolaridad, no hay grandes diferencias entre los porcentuales de despidos en relación al sexo y el color, y la construcción civil fue el sector económico con mayor tasa de escolaridad los apagones sin justa causa entre los demás. Por último, se constató que el modelo logístico generado tuvo un ajuste razonable a los datos, con exactitud obtenida por la curva ROC del 70%, y sus variables explicativas “sexo”, “color”, “franja etaria”, “grado de escolaridad”, “portador de deficiencia”, “alejamiento del trabajo”, “tiempo de empleo”, “renta media en salarios mínimos”, “sector económico” y “tamaño del establecimiento en número de empleados” fueron significativas al nivel del 1%.

Palabras-clave

Seguro de personas. Pérdida de la renta. RAIS. Modelos Lineales Generalizados. Regresión Logística.

Sumario

1. Introdução

Perder o emprego nos dias atuais é uma das maiores preocupações dos brasileiros. De acordo com o indicador elaborado pela Confederação Nacional da Indústria (CNI), em junho de 2018, o Índice de Medo do Desemprego (IMD) foi de 67,9 pontos, em uma escala que vai de 0 a 100, atingindo o maior valor da série histórica desde 1996.

Segundo dados de estudo trimestral realizado pelo Instituto de Pesquisa Econômica Aplicada (IPEA) divulgado em 12 de abril de 2018, a taxa de desemprego da população brasileira nos primeiros três meses daquele ano era de 13,1%, mantendo-se estável em torno de 12,5% quando considerados os ajustes devidos ao efeito sazonal. Essa desocupação esteve mais acentuada na Região Nordeste (15,8%) e foi menor na Região Sul (8,4%). Isso ajuda a explicar por que o IMD no Nordeste foi o mais alto dentre as demais regiões, apresentando 74,1 pontos em junho de 2018, conforme publicação da CNI.

O fantasma do desemprego afeta mais os jovens até 24 anos, cujo IMD de 70,8 pontos pode estar correlacionado com o fato de este ser o público com a maior taxa de desocupação (28,1%). Segundo a análise do IPEA, essa parcela da população possui as menores chances de sair do desemprego e maiores probabilidades de entrar na desocupação. No entanto, esse temor pelo desemprego também é alto nas pessoas das faixas etárias de 25 a 34 anos, 35 a 44, 45 a 54 e acima de 55 anos, com o IMD em torno de 67 pontos.

Toda essa preocupação do cidadão brasileiro com a possibilidade de ficar desempregado tem demandado a procura por modalidades de seguros abrangendo cobertura do pagamento de contas no caso da falta da renda, como, por exemplo, mensalidades de escolas, parcelas de condomínios e aluguéis, conforme mostrou a reportagem “Seguro para pagar contas cresce na crise”, do jornal O Globo, publicada em 09 de junho de 2018 e assinada por Daiane Costa[1].

O tema ganha especial atenção em um momento no qual o Governo claramente se posiciona pela diminuição do Estado. Destacamos recente fala da Sra. Solange Vieira, da Superintendência de Seguros Privados (Susep), que faz parte do trecho de notícia veiculada pela CNseg: “Considerando o SUS, o INSS e o seguro desemprego, a participação do seguro público no seguro é maior que a do privado. ‘Temos de rever isso, trazendo para a iniciativa privada proteções como o seguro desemprego’”, disse a superintendente da autarquia na CONSEGURO 2019 (CNSEG, 2019)[2].

[1] COSTA, Daiane. Seguro para pagar contas cresce na crise. **O Globo**. 9 mai. 2018. Disponível em: https://oglobo.globo.com/economia/defesa-do-consumidor/seguro-para-pagar-contas-cresce-na-crise-22666167. Acesso em: 12 set. 2018.

[2] CNSEG. Susep estuda trazer o seguro desemprego para a iniciativa privada. 4 set. 2019. **Site**. Disponível em: http://cnseg.org.br/noticias/susep-estuda-trazer-o-seguro-desemprego-para-a-iniciativa-privada.html. Acesso em: 4 set. 2019.

Muitos desses produtos são atrelados ao seguro de vida, sendo a cobertura da perda de renda um adicional na apólice. Tendo em vista que a precificação de qualquer modalidade de seguros necessita da apuração do respectivo risco associado ao evento, é necessário um estudo para mensurar esse valor.

O presente estudo tem como fundamento propor um modelo de regressão logístico para estimar o risco de uma pessoa que trabalha no regime CLT e com contrato por tempo indeterminado ser demitida sem justa causa, além da utilização dessa estimativa na precificação de um seguro desemprego de curto prazo. Para alcançar esse objetivo utilizamos uma base de dados pública, logo, este trabalho poderia ser de relevância para pequenas seguradoras sem massa de dados suficiente para a precificação desse tipo de seguro com crescente demanda. O restante deste texto está organizado da seguinte forma: o segundo capítulo apresenta o desenvolvimento do modelo e, no terceiro, estão os resultados. A conclusão é feita no último capítulo.

2. Desenvolvimento do modelo logístico

2.1 Base de dados

Existe uma grande carência de dados para modelagem do risco de desemprego, primeiro, por envolver relações particulares e, segundo, por ser um mercado pouco explorado. Este é um especial dilema para pequenas empresas ou seguradoras que estejam iniciando a operação, e por isso há dificuldades em utilizar seus dados para modelagem estatística, pois estes não existem ou não são confiáveis. Nesse caso, uma opção matematicamente viável é fazer uso da Teoria da Credibilidade[3], mesclando experiências históricas de dados externos à seguradora com aqueles de suas experiências de riscos mais recentes. Daí aplica-se tal conceito na precificação do seguro.

Devido aos problemas relatados acima, sobre a falta de dados para o estudo desejado, recorreu-se à busca por microdados oriundos de órgãos públicos do Governo Federal, uma base oficial e confiável de informação.

O antigo Ministério do Trabalho (MTb), que atualmente faz parte do Ministério da Economia, disponibilizava anualmente, até o final de 2018, microdados gerados a partir da Relação Anual de Informações Sociais (RAIS), uma das principais fontes de informação sobre o mercado de trabalho formal brasileiro. É utilizada constantemente pelo Governo para elaborar e acompanhar políticas públicas de trabalho, emprego e renda.

A RAIS foi definida pelo MTb como *"um cadastro administrativo, instituído pelo Decreto nº 76.900, de 23/12/1975, de âmbito nacional, periodicidade anual e de declaração obrigatória para todos os estabelecimentos do setor público e privado, inclusive para aqueles que não registraram vínculos empregatícios no exercício"* (grifo nosso) (SNIF, 2016, p.5).

[3] A Teoria da Credibilidade é uma técnica que consiste em mesclar a experiência mais recente da seguradora com a de riscos similares de períodos anteriores de outras empresas do ramo ou do mercado como um todo. É muito importante para os casos em que a empresa possui poucos dados de sinistros (FERREIRA, 2002).

Esse cadastro possui como principais variáveis o estoque de empregos posicionados em 31 de dezembro, segundo gênero, faixa etária, grau de escolaridade, tempo de serviço e rendimentos, com dados desagregados em nível ocupacional, geográfico e setorial, contemplando aproximadamente 99% do universo do mercado formal brasileiro, sendo fornecidos por estabelecimento e por trabalhador.

Tendo-se em vista a relevância e a elevada quantidade de informação sobre o mercado de trabalho formal no Brasil oferecida por um órgão oficial e especializado do Governo Federal, optou-se por utilizar os microdados da RAIS trabalhador neste estudo.

A base de dados escolhida para a aplicação do modelo logístico foi gerada em 15 de setembro de 2018, a partir dos microdados da RAIS provenientes do município do Rio de Janeiro[4] no ano de 2016. Estes, à época, eram os dados mais recentes disponíveis[5].

Como o foco do estudo são os empregados maiores de 18 anos, regidos pela CLT e com contrato de trabalho por tempo indeterminado, foram excluídos da base de dados os registros que não se enquadravam em tal perfil.

A variável resposta (Y) a ser utilizada no modelo logístico[6] será uma dummy, onde:

$$Y = \begin{cases} 1, \text{ para empregados com rescisão sem justa causa por iniciativa do empregador} \\ 0, \text{ caso contrário, isto é, empregados que permaneceram ativos ao longo de 2016} \end{cases}$$

Logo, a base de dados resultante de todas as exclusões efetuadas passou a ser de 2.390.246 registros, correspondentes a trabalhadores maiores de 18 anos, regidos pela CLT, com contrato de trabalho por tempo indeterminado em estabelecimento localizado no município do Rio de Janeiro. Destes, 1.851.661 (77,5%) permaneceram ativos ao longo de todo ano de 2016, e 538.585 (22,5%) foram desligados sem justa causa.

[4] Optou-se pelo município do Rio de Janeiro por ser uma cidade com grande número de empregados formais e por ter um mercado segurador atuante.

[5] Fonte: BRASIL. Secretaria de Trabalho. PDET. Microdados RAIS e CAGED. **Site Ministério da Economia**. 19 mai. 2016. Disponível em: http://pdet.mte.gov.br/microdados-rais-e-caged. Acesso em: 12 jan. 2020.

[6] O software utilizado para a análise estatística dos dados e para a elaboração do modelo logístico foi o RStudio (Version 1.1.383), com a versão 3.4.3 do R.

2.2 Análise exploratória dos dados

A Tabela 1 revela a distribuição do número de empregados em relação ao perfil e à situação no emprego. Quando a análise do percentual de demitidos é feita considerando-se o gênero dos empregados, verifica-se que 24% dos homens que trabalharam em 2016 com carteira assinada foram desligados sem justa causa. No caso das mulheres, o equivalente se dá com 20%.

Em relação ao nível de escolaridade, os trabalhadores com baixa instrução estão entre os que possuem as maiores taxas de demissões, ou seja, quanto menor a escolaridade, maior o índice de perda do emprego. Pode-se notar, ainda, que, dentre os empregados com mestrado ou doutorado, menos de 10% foram despedidos.

Os trabalhadores indígenas e amarelos foram os que apresentaram os maiores percentuais de demissões, em torno de 25%, enquanto os de cor branca foram os únicos abaixo da média, com 20%. No entanto, percebe-se que esse dado pode não representar a realidade, uma vez que 6% dos empregados da base de dados estão sem cor identificada.

Do total de empregados com menos de 30 anos de idade, mais de 26% foram desligados ao longo de 2016. Os jovens entre 18 e 24 anos possuem taxa em torno de 28%, ou seja, 6 pontos percentuais acima da média total de demitidos. Vê-se que essa taxa se reduz à medida que a idade do empregado aumenta, tendo os públicos de 50 a 64 anos (17,5%) e de 65 e mais anos (16,7%) apresentado os menores percentuais de demissões sem justa causa.

Os microdados da RAIS também contêm informações sobre trabalhadores com deficiência. Sendo assim, a Tabela 1 permite observar que, dos empregados portadores de deficiência, cerca de 14% tiveram seu contrato encerrado sem justa causa. Contudo, aproximadamente 86% continuaram em atividade nas empresas.

Tabela 1 – Perfil social dos empregados em relação à situação do emprego

Variáveis		Situação				Total	
		Demitido		Ativo			
		Qtde	%	Qtde	%	Qtde	%
Gênero	Masculino	344.954	24,0	1.090.729	76,0	1.435.683	100,0
	Feminino	193.631	20,3	760.932	79,7	954.563	100,0
Escolaridade	Sem escolaridade	72.920	26,7	200.455	73,3	273.375	100,0
	Fundamental	119.556	25,7	345.525	74,3	465.081	100,0
	Médio	275.572	22,8	930.437	77,2	1.206.009	100,0
	Superior	68.968	16,1	360.346	83,9	429.314	100,0
	Mestrado	1.180	9,5	11.284	90,5	12.464	100,0
	Doutorado	389	9,7	3.614	90,3	4.003	100,0
Cor/Raça	Indígena	1.546	25,7	4.467	74,3	6.013	100,0
	Branca	219.564	20,3	860.299	79,7	1.079.863	100,0
	Preta	53.290	23,9	169.862	76,1	223.152	100,0
	Amarela	4.112	25,5	12.010	74,5	16.122	100,0
	Parda	224.471	24,5	690.021	75,5	914.492	100,0
	Não identificada	35.602	23,6	115.002	76,4	150.604	100,0
Faixa etária	18 a 24 anos	85.474	27,8	222.273	72,2	307.747	100,0
	25 a 29 anos	98.414	26,3	276.276	73,7	374.690	100,0
	30 a 39 anos	176.177	23,1	585.621	76,9	761.798	100,0
	40 a 49 anos	102.061	20,1	406.398	79,9	508.459	100,0
	50 a 64 anos	69.968	17,5	328.769	82,5	398.737	100,0
	65 anos ou mais	6.491	16,7	32.324	83,3	38.815	100,0
Deficiência	Possui deficiência	3.103	14,1	18.837	85,9	21.940	100,0
	Não possui deficiência	535.482	22,6	1.832.824	77,4	2.368.306	100,0

Fonte: Elaboração própria.

As informações referentes às características da vida laboral dos empregados estão representadas mais adiante neste texto, na Tabela 2.

Os setores de atuação dos estabelecimentos vinculados aos empregados puderam ser extraídos dos microdados da RAIS através do campo "subsetor IBGE". Essa variável possui 25 categorias de atividades econômicas, das quais 13 se referem ao setor industrial.

Visando a uma melhor apresentação dos dados, as atividades industriais foram agregadas em “indústria”, e as de comércio varejista e comércio atacadista, em “comércio”.

Os setores “serviços industriais de utilidade pública”, “instituições de crédito, seguros e capitalização”, “administração pública direta e autárquica” e “agricultura, silvicultura, criação de animais, extrativismo vegetal” apresentaram poucas frequências de empregados, sendo assim incorporados no campo “outros setores”.

A Tabela 2 revela que o setor de construção civil demitiu quase a metade do total de funcionários em 2016, com uma taxa de desligamentos de 48%. O setor imobiliário e o de alojamento, alimentação e manutenção demitiram 23,2% do seu quadro de funcionários. O comércio e a indústria também tiveram taxas de demissões acima de 20%. Os estabelecimentos de outros setores econômicos demitiram menos de 10% do total de trabalhadores.

Em relação ao tempo de emprego, percebe-se que os empregados com menos de um ano na empresa estão mais sujeitos a demissões do que aqueles mais antigos. Dos trabalhadores com mais de 10 anos de empresa, em torno de 6% foram demitidos. Para aqueles com menos de um ano, esse percentual é quase cinco vezes maior.

Com intuito de facilitar a análise descritiva da remuneração média em salários mínimos dos empregados, as classes de frequência da variável original foram agregadas em novas faixas, uma vez que várias categorias apresentavam baixas frequências. Além disso, constatou-se que havia na base de dados 70.972 casos sem informações para a remuneração média em salários mínimos dos empregados, representando cerca de 3% do total.

Dentre os empregados que receberam até um salário mínimo em 2016, mais da metade foi demitida. Dos trabalhadores com remuneração média maior do que um salário mínimo e menor do que dois salários mínimos, 26% foram desligados. A classe de frequência com o menor percentual de desligamentos foi a que abrangeu os empregados que receberam mais de 10 salários mínimos, com 13,7% de demissões.

De acordo com a Tabela 2, os trabalhadores cuja empresa possui menos de 20 funcionários são os que apresentaram maiores percentuais de demissões (27,3%). Já os trabalhadores das empresas com 500 e mais funcionários foram demonstraram ter as menores taxas (16,4%).

Deve-se notar que há empregados que, algumas vezes, seja involuntariamente ou por vontade própria, necessitam se ausentar do trabalho por determinado motivo, como licença sem vencimentos, licenças médicas, licença-maternidade, acidente de trabalho etc.

Os trabalhadores afastados em 2016 demitidos sem justa causa somaram 13,8%, sendo de 23,3% o índice de demissão entre os que não se licenciaram.

Tabela 2 – Características do trabalho dos empregados em relação à situação do emprego

Variáveis		Situação				Total	
		Demitido		Ativo			
		Qtde	%	Qtde	%	Qtde	%
Setor econômico	Outros setores	11.481	9,5	109.472	90,5	120.953	100,0
	Indústria	41.894	20,6	161.841	79,4	203.735	100,0
	Construção civil	92.543	48,0	100.187	52,0	192.730	100,0
	Comércio	110.514	21,6	400.238	78,4	510.752	100,0
	Imobiliário	116.975	23,2	387.106	76,8	504.081	100,0
	Transportes e comunicações	37.195	18,5	163.320	81,5	200.515	100,0
	Serviços de alojamento/ alimentação/manutenção	93.028	23,2	308.262	76,8	401.290	100,0
	Serviços médicos/ odontológicos/veterinários	19.670	14,3	117.911	85,7	137.581	100,0
	Ensino	15.285	12,9	103.324	87,1	118.609	100,0
Tempo de emprego[1]	Menos de 1 ano	221.190	29,4	531.376	70,6	752.566	100,0
	De 1 a 4 anos	258.289	24,2	808.661	75,8	1.066.950	100,0
	De 5 a 9 anos	41.888	13,1	277.112	86,9	319.000	100,0
	10 e mais anos	16.830	6,7	233.707	93,3	250.537	100,0
Remuneração média em salários mínimos (SM)[2]	Até 1 SM	40.557	51,9	37.630	48,1	78.187	100,0
	Mais de 1 até 1,5 SM	195.482	26,0	557.326	74,0	752.808	100,0
	Mais de 1,5 até 2 SM	99.321	20,8	378.239	79,2	477.560	100,0
	Mais de 2 até 3 SM	84.993	19,9	342.238	80,1	427.231	100,0
	Mais de 3 até 5 SM	48.536	17,3	232.118	82,7	280.654	100,0
	Mais de 5 até 10 SM	27.345	15,4	149.903	84,6	177.248	100,0
	Mais de 10 SM	17.223	13,7	108.363	86,3	125.586	100,0
Tamanho do estabelecimento	Menos de 20 funcionários	189.713	27,3	506.098	72,7	695.811	100,0
	De 20 a 99 funcionários	137.437	22,4	475.781	77,6	613.218	100,0
	De 100 a 499 funcionários	112.985	23,4	369.505	76,6	482.490	100,0
	500 e mais funcionários	98.450	16,4	500.277	83,6	598.727	100,0
Afastamentos	Ficou afastado	25.796	13,8	161.218	86,2	187.014	100,0
	Não ficou afastado	512.789	23,3	1.690.443	76,7	2.203.232	100,0

Fonte: Elaboração própria.
Nota: (1) Sem registro de tempo de emprego para 1.193 empregados.
(2) Sem registro de remuneração média para 70.972 empregados.

3. Resultados do modelo logístico

O modelo logístico aplicado levou em consideração os aspectos teóricos apresentados por Jong e Heller (2008). Primeiramente, a base de dados foi segregada em duas: uma parte para treino do modelo e outra para testá-lo, ambas selecionadas pelo método da amostragem aleatória simples[7].

Na amostra de treino foram estimados os parâmetros do modelo com base na maximização do *log* da função de verossimilhança e realizados os testes de Wald e da razão de verossimilhança, para verificar se os mesmos eram significativos. Em seguida, foi comparado o modelo gerado com outros de menor número de variáveis explicativas através do algoritmo *Stepwise*[8]. Por último, nos dados da amostra de teste, examinou-se a qualidade do modelo escolhido por meio da análise da curva ROC[9].

Foi necessário excluir da base de dados todos os empregados com ausência de informação em pelo menos uma variável. Com isso, registros de 2.170.774 trabalhadores foram utilizados para a elaboração do modelo logístico.

A existência de variáveis com frequências reduzidas em algumas categorias pode causar problemas no ajuste do modelo logístico. Dessa forma, decidiu-se recategorizar as variáveis "cor", "faixa etária" e "escolaridade do empregado", que apresentaram categorias percentuais inferiores a 1,6%. Assim, foi criada uma nova variável para "cor" do empregado com apenas duas categorias: "branca/amarela" e "preta/parda/indígena". Na escolaridade, os níveis "mestrado" e "doutorado" foram agregados ao nível superior. Por último, a faixa etária de 65 e mais anos passou a fazer parte do grupo com 50 a 64 anos, sendo agora classificada como "50 e mais anos".

Visando a minimizar o risco de *overfitting*[10], a base de dados foi dividida em duas partes: 70% das observações foram destinadas ao treino do modelo (1.519.541 empregados) e 30%, para o teste do ajuste (651.233 empregados). A seleção ocorreu pelo método da amostragem aleatória simples.

---

[7] Método de amostragem em que todos os elementos possuem a mesma probabilidade de serem sorteados.

[8] *Stepwise* é um algoritmo que possibilita a escolha automatizada de modelos em relação às variáveis explicativas a serem utilizadas.

[9] Curva elaborada a partir de todos os pontos de corte possíveis em relação à sensibilidade (proporção de eventos classificados corretamente como ocorridos) e à especificidade (proporção de eventos classificados corretamente como não ocorridos).

[10] *Overfitting* ou sobreajuste é o caso em que o modelo gerado só está bem ajustado aos dados atuais, ou seja, aqueles utilizados na construção do modelo. Quando isso ocorre, este é ineficiente para a predição de novos dados.

Primeiramente, foi elaborado um modelo logístico a partir dos dados de treino, tendo como variável-resposta a demissão ou não do empregado, acrescida das seguintes variáveis explicativas: sexo, escolaridade, portador de deficiência, cor, faixa etária, afastamento do empregado, tempo de empresa, setor econômico do estabelecimento, tamanho da empresa e remuneração média em salários.

A Tabela 3 a seguir mostra o resultado do teste de razão de verossimilhança, que checa se todos os parâmetros estimados do modelo são estatisticamente iguais a zero ($\beta_1 = \beta_2 = ... = \beta_p = 0$), a partir da comparação entre o modelo ajustado e aquele apenas com o intercepto.

A estatística para o teste de razão de verossimilhança foi calculada com base na distribuição de qui-quadrado, cujo valor encontrado foi de 143.866 (*Deviance*). Nota-se pelo *p-valor* que a hipótese nula foi rejeitada, ou seja, existe pelo menos um $\beta \neq 0$.

Tabela 3 – Teste da Razão de Verossimilhança

	Resid Df	Resid Dev	Df	Deviance	Pr(>Chi)
Modelo só com intercepto	1519540	1603710			
Modelo completo	1519509	1459845	31	143866	0

Fonte: Elaboração própria.

De forma similar, pode-se aplicar a lógica acima, utilizando-se novamente a distribuição de qui-quadrado para testar se todas as variáveis explicativas no modelo foram significativas ao nível de 1%, sendo tal hipótese confirmada na Tabela 4:

Tabela 4 – Teste da Razão de Verossimilhança para as variáveis

	Df	Deviance	Resid Df	Resid Dev	Pr(>Chi)
NULL			1519540	1603710	0
Sexo	1	3436	1519539	1600274	0
Escolaridade	3	11309	1519536	1588966	0
Faixa etária	4	9600	1519532	1579366	0
Cor	1	1143	1519531	1578223	0
Portador de deficiência	1	464	1519530	1577760	0
Tempo de emprego	3	42273	1519527	1535487	0
Remuneração média	6	22836	1519521	1512650	0
Tamanho do estabelecimento	3	9015	1519518	1503636	0
Afastamentos em 2016	1	480	1519517	1503155	0
Setor econômico	8	43311	1519509	1459845	0

Fonte: Elaboração própria.

A Tabela 5 apresenta as estimativas de máxima verossimilhança dos parâmetros do modelo para cada covariável, além de seus respectivos erros padrão e da *odds ratio*. Também traz os valores observados da estatística de Wald e seus *p-valores*, que foram utilizados para testar cada $\beta_i = 0$. Concluiu-se que todas foram significativas ao nível de 1%.

O nível base do modelo está representado por “sexo: masculino”, “sem escolaridade”, “faixa etária: 18 a 24 anos”, “cor: branca/amarela”, “portador de deficiência”, “tempo de emprego: menos de 1 ano”, “remuneração: até 1 SM”, “tamanho da empresa: menos de 20 funcionários”, “afastamentos: ficou afastado” e “setor econômico do estabelecimento: outros setores”.

Dentre os níveis de maior impacto do modelo, destacam-se “setor econômico: construção civil” (1,7806), “tempo de emprego: 10 anos e mais” (-1,4335), “tamanho do estabelecimento: 500 e mais funcionários” (-0,5736) e “remuneração média: mais de 2 até 3 SM” (-1,6945).

Se considerarmos a razão de chances (*odds ratio*) das covariáveis acima e supondo que esse seja o modelo escolhido, é possível afirmar que uma pessoa trabalhando na construção civil possua 5,9 vezes mais chances de perder o emprego frente aos demais setores. No entanto, para os empregados que com remuneração média de dois a três salários mínimos, a chance da demissão é bem menos provável do que para os demais (0,1837).

Tabela 5 – Estimativas dos parâmetros do modelo logístico

Coefficients:		Estimate	Std Error	z value	Pr(>|z|)	odds ratio
	Intercepto (β_0)	-0,4200	0,0324	-13,0	0,0000	0,6571
Sexo:	Feminino (β_1)	-0,1020	0,0046	-22,0	0,0000	0,9030
Escolaridade:	Fundamental (β_2)	-0,0674	0,0074	-9,1	0,0000	0,9349
Escolaridade:	Médio (β_3)	-0,1526	0,0068	-22,4	0,0000	0,8585
Escolaridade:	Superior (β_4)	-0,2902	0,0096	-30,4	0,0000	0,7481
Faixa etária:	25 a 29 anos (β_5)	0,1269	0,0073	17,3	0,0000	1,1353
Faixa etária:	30 a 39 anos (β_6)	0,1154	0,0067	17,3	0,0000	1,1224
Faixa etária:	40 a 49 anos (β_7)	0,0410	0,0074	5,6	0,0000	1,0419
Faixa etária:	50 anos ou mais (β_8)	0,0245	0,0080	3,1	0,0022	1,0248
Cor:	Indígena/Preta/Parda (β_9)	0,0755	0,0043	17,5	0,0000	1,0784
Deficiência:	Não possui deficiência (β_{10})	0,3552	0,0256	13,8	0,0000	1,4265
Tempo de emprego:	De 1 a 4 anos (β_{11})	-0,1416	0,0046	-31,1	0,0000	0,8679
Tempo de emprego:	De 5 a 9 anos (β_{12})	-0,7780	0,0078	-99,5	0,0000	0,4593
Tempo de emprego:	10 e mais anos (β_{13})	-1,4335	0,0115	-124,2	0,0000	0,2385
Remuneração média:	Mais de 1 até 1,5 SM (β_{14})	-1,2832	0,0100	-127,7	0,0000	0,2771
Remuneração média:	Mais de 1,5 até 2 SM (β_{15})	-1,6344	0,0106	-153,7	0,0000	0,1951
Remuneração média:	Mais de 2 até 3 SM (β_{16})	-1,6945	0,0109	-154,8	0,0000	0,1837
Remuneração média:	Mais de 3 até 5 SM (β_{17})	-1,6472	0,0118	-139,7	0,0000	0,1926
Remuneração média:	Mais de 5 até 10 SM (β_{18})	-1,5656	0,0134	-116,8	0,0000	0,2090
Remuneração média:	Mais de 10 SM (β_{19})	-1,4972	0,0155	-96,7	0,0000	0,2238
Tamanho do estabelecimento:	De 20 a 99 funcionários (β_{20})	-0,3030	0,0055	-55,4	0,0000	0,7386
Tamanho do estabelecimento:	De 100 a 499 funcionários (β_{21})	-0,2348	0,0060	-39,0	0,0000	0,7908
Tamanho do estabelecimento:	500 e mais funcionários (β_{22})	-0,5736	0,0062	-92,0	0,0000	0,5635
Afastamento:	Não ficou afastado (β_{23})	0,2316	0,0097	24,0	0,0000	1,2606
Setor econômico:	Indústria (β_{24})	0,6610	0,0151	43,7	0,0000	1,9367
Setor econômico:	Construção civil (β_{25})	1,7806	0,0149	119,8	0,0000	5,9335
Setor econômico:	Comércio (β_{26})	0,3460	0,0143	24,2	0,0000	1,4134
Setor econômico:	Imobiliário (β_{27})	0,7192	0,0141	51,2	0,0000	2,0529
Setor econômico:	Transportes e comunicações (β_{28})	0,5129	0,0153	33,5	0,0000	1,6701
Setor econômico:	Serviços de alojamento/ alimentação/manut. (β_{29})	0,4916	0,0144	34,1	0,0000	1,6349
Setor econômico:	Serviços médicos/odontológicos/ veterinários (β_{30})	0,2209	0,0167	13,2	0,0000	1,2472
Setor econômico:	Ensino (β_{31})	-0,1998	0,0181	-11,0	0,0000	0,8189

Fonte: Elaboração própria.

Ao estimar os parâmetros do modelo, o Software R calcula as medidas do *Deviance* do modelo nulo (1603710, com 1519540 graus de liberdade) e o do modelo ajustado (1459845, com 1519509 graus de liberdade), além do valor do AIC[11] (1459909).

É recomendável verificar outros modelos com menos variáveis explicativas. Isso pode ser feito por meio do algoritmo *Stepwise*, que possibilita selecionar modelos através do AIC.

Utilizando-se a função *stepAIC* do pacote MASS do R no modelo gerado com todas as variáveis, nota-se, na Tabela 6, que o modelo atual é o mais adequado, uma vez que os valores do AIC e do *Deviance* vão aumentando conforme variáveis são excluídas.

Tabela 6 – Resultado do Algoritmo Stepwise

	Df	Deviance	AIC
<none>		1459845	1459909
– Portador de deficiência	1	1460050	1460112
– Cor	1	1460152	1460214
– Sexo	1	1460330	1460392
– Faixa etária	4	1460402	1460458
– Afastamentos em 2016	1	1460440	1460502
– Escolaridade	3	1460936	1460994
– Tamanho do estabelecimento	3	1468848	1468906
– Tempo de emprego	3	1485468	1485526
– Remuneração média em SM	6	1487667	1487719
– Setor econômico	8	1503155	1503203

Fonte: Elaboração própria.

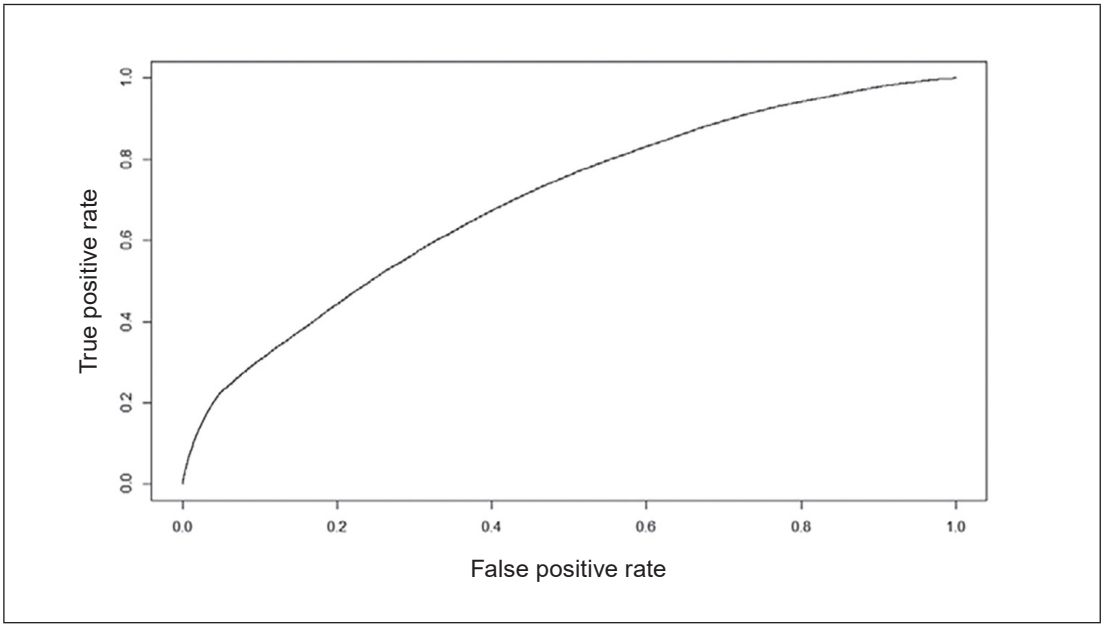
A curva ROC foi a ferramenta utilizada para verificar a qualidade do modelo gerado. A acurácia é encontrada pela área abaixo da curva, tendo como base as taxas de verdadeiro positivo e falso positivo.

Conforme explicado anteriormente, construiu-se a curva ROC a partir dos dados separados para o teste do modelo, tentando-se assim contornar o problema do *overfitting*.

[11] *Akaike's Information Criterion* (AIC) é um critério de seleção de modelos baseado em uma medida que equilibra o ajuste do modelo com o número de parâmetros. Sua fórmula é dada por AIC = -2L + 2P, onde L é o logaritmo da função de verossimilhança do modelo e *P* **é o número de parâmetros. Recomenda-se escolher o modelo que possuir o menor valor do AIC** (JONG; HELLER, 2008).

Observando-se o Gráfico 1, os dados parecem estar razoavelmente ajustados. O valor da acurácia é de 70%.

Gráfico 1 – Curva ROC



Como o modelo se mostrou viável aos dados, já é possível realizar previsões utilizando-se tanto a *odds ratio* quanto as probabilidades de ocorrência dos eventos. Para tal, basta que sejam consideradas as variáveis da Tabela 5 e que seja aplicada a fórmula abaixo, na qual X são as variáveis explicativas, e β, os parâmetros estimados:

$$P(Y = 1|X_1, X_2, \cdots, X_p) = \pi = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}$$

Seja:

$$W = \beta^T \times X$$

Logo, a fórmula para se calcular o risco de perda de renda de um empregado celetista, com mais de 18 anos e com contrato por tempo indeterminado, é dada por:

$$\pi = \frac{e^{W}}{1 + e^{W}}$$

De acordo com a fórmula acima, deduz-se que o perfil de maior risco corresponde aos empregados do sexo feminino, com nível fundamental completo, na faixa etária de 25 a 29 anos, de etnia indígena, negra ou parda. Tais pessoas não são portadoras de deficiência e seu tempo de emprego é de 1 a 4 anos. Recebem até um salário mínimo, trabalhando em empresas de 100 até 499 funcionários. Não tiveram quaisquer afastamentos e atuam no setor de construção civil. A probabilidade calculada para a perda do emprego desse perfil é de 83,3%.

Como uma simples aplicação, podemos definir um seguro desemprego contratado com cobertura de um ano contra a perda da renda do perfil acima, capaz de pagar uma indenização de R$1.500,00 durante seis meses em caso de sinistro e que considere uma condição de mercado de taxa de juros de 6% a.a. (0,487% a.m.). Logo, o valor do prêmio puro mensal a ser pago postecipadamente durante um ano pode ser calculado da seguinte forma:

$$1500\, P_{puro} = 1500 \times 0{,}0518 \times 5{,}89 = \text{R\$ } 458{,}30$$

O valor de 0,0518 representa o risco mensal do perfil associado e 5,899 é uma anuidade financeira.

Aplicando-se o mesmo conceito exposto acima, é viável também calcular o prêmio puro mensal para o perfil de menor risco (probabilidade mensal de 0,132%) e para um perfil de risco médio (probabilidade mensal de 3,44%), obtendo-se, respectivamente, R$ 11,71 e R$ 304,09.

Claramente observa-se que, devido ao risco, o custo do seguro seria muito elevado para determinadas faixas. Com isso, a empresa teria que focar seus esforços nas faixas viáveis de comercialização.

4. Conclusão

Além da elaboração do modelo de regressão logística para estimar o risco da perda da renda de um empregado celetista, maior de 18 anos e com contrato de trabalho por tempo indeterminado, bem como a consequente utilização na tarifação do seguro, o presente estudo também possibilitou a verificação do perfil do trabalhador em relação às demissões sem justa causa ocorridas ao longo de 2016. Notou-se que o percentual de perda de emprego diminuía conforme o grau de escolaridade aumentava, não houve grandes diferenças entre as taxas de demissões por gênero e raça e a área de construção civil foi a mais afetada em relação aos desligamentos sem justa causa.

Embora o modelo logístico produzido tenha se mostrado aceitável em relação às variáveis utilizadas, em virtude dos dados estarem disponíveis como um retrato do ano não foi possível considerar informações macroeconômicas diretamente ligadas ao desemprego, como o PIB, a inflação e a taxa de juros, entre outras.

No entanto, esse modelo pode ser aplicado em uma empresa que realiza uma tarifação mensal e lança mão de dados mais recentes para a mensuração de um seguro de curto prazo, sendo este ajustado rapidamente em caso de mudanças do cenário.

Como sugestão para novos trabalhos, seria indicado um estudo comparativo entre os resultados obtidos para previsão de sinistros pelo modelo de regressão logística e aqueles gerados a partir de outras técnicas, como as redes neurais e as árvores de decisão.

5. Referências bibliográficas

BRASIL. Secretaria de Trabalho. PDET. Microdados RAIS e CAGED. Site Ministério da Economia. 19 mai. 2016. Disponível em: http://pdet.mte.gov.br/microdados-rais-e-caged. Acesso em: 12 jan. 2020.

BOLFARINE, H.; BUSSAB, W. O. **Elementos de amostragem**. 1. ed. São Paulo: Blucher, 2005.

BORGES, A. Impactos do desemprego e da precarização sobre famílias metropolitanas. **Revista Brasileira de Estudos Populacionais**, São Paulo, v.23, n.2, p.205-222, jul./dez. 2006. Disponível em: http://www.scielo.br/pdf/rbepop/v23n2/a02v23n2. Acesso em: 29 out. 2018.

BÜHLMANN, H.; GISLER, A. **A Course in Credibility Theory and its Applications**. 1. ed. The Netherlands: Springer, 2005.

CNI. Confederação Nacional da Indústria. Indicadores CNI – Julho/2018 – nº 2. Disponível em : http://www.portaldaindustria.com.br/estatisticas/medo-do-desemprego-satisfacao-com-a-vida/. Acesso em: 9 set. 2018.

CNSEG. Susep estuda trazer o seguro desemprego para a iniciativa privada. 4 set. 2019. **Site**. Disponível em: http://cnseg.org.br/noticias/susep-estuda-trazer-o-seguro-desemprego-para-a-iniciativa-privada.html. Acesso em: 4 set. 2019.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos lineares generalizados e extensões**. 1. ed. São Paulo: USP, 2008.

COSTA, Daiane. Seguro para pagar contas cresce na crise. **O Globo**. 9 mai. 2018. Disponível em: https://oglobo.globo.com/economia/defesa-do-consumidor/seguro-para-pagar-contas-cresce-na-crise-22666167. Acesso em: 12 set. 2018.

DOBSON, A. **An introduction to generalized linear models**. 2. ed. Londres: Chapman & Hall/CRC, 2002.

FÁVERO, L. P. *et al*. **Análise de dados:** modelagem multivariada para tomada de decisões. 1. ed. Rio de Janeiro: Elsevier, 2009.

FERREIRA. P. P. **Modelos de precificação e ruína para seguros de curto prazo**. 1. ed. Rio de Janeiro: FUNENSEG, 2002.

FREITAS, M. A. L. de. Modelo logístico aplicado ao mercado de seguros de auto no Brasil: cálculo da probabilidade de sinistros. **Revista Fee**, Porto Alegre, v. 37, n. 3. 2009. Disponível em: https://revistas.fee.tche.br/index.php/indicadores/article/view/2333. Acesso em: 9 set. 2018.

GARCIA, F. T. *et al*. Proposta de um modelo probabilístico de risco de crédito com a aplicação da técnica de regressão logística. **Revista Gestão & Conhecimento**, v. 7, n.1, p. 175-207, jan./jun. 2013. Disponível em: https://www.facet.br/gc/artigos/resumo.php?artigo=55. Acesso em: 2 nov. 2018.

GONÇALVES, J. R. **Relação entre desemprego e aversão ao risco**: uma análise do mercado de trabalho de Fortaleza-CE. 2014. 34 f. Dissertação (Mestrado em Economia) – Universidade Federal do Ceará, Fortaleza. Disponível em: http://www.repositorio.ufc.br/bitstream/riufc/9989/1/2014_dissert_jrgoncalves.pdf. Acesso em: 29 out. 2018.

GUJARATI, D. **Econometria básica**. 4. ed. Rio de Janeiro: Elsevier, 2006.

IPEA. Instituto de Pesquisa Econômica Aplicada. **Carta de Conjuntura 2018**. 2º trimestre, n. 39. Disponível em: http://www.ipea.gov.br/portal/index.php?option=com_content&view=article&id=32921. Acesso em: 9 set. 2018.

JONG, P. D.; HELLER, G. Z. **Generalized linear models for insurance data**. 1. ed. Cambridge: Cambridge University Press, 2008.

KAAS, R.; GOOVAERTS, M.; DHAENE, J.; DENUIT, M. **Modern actuarial risk theory**: using R. 2.ed.S.l: Springer, 2008.

MCCULLAGH, P.; NELDER, J.A. **Generalized linear models**. 2. ed. Londres: Chapman & Hall, 1989.

PAULA. G. A. **Modelos de regressão com apoio computacional**. 1. ed. São Paulo: USP, 2004.

RODRIGUES, W.; PARREIRA, L.A. Análise do risco ao desemprego entre grupos demográficos no município de Palmas – TO: uma aplicação do modelo de regressão logístico binomial. **Informe Gepec**, Toledo, v. 17, n. 1, p. 23-33, jan./jun. 2013. Disponível em: http://e-revista.unioeste.br/index.php/gepec/article/view/5905. Acesso em: 29 out. 2018.

SNIF. Sistema Nacional de Informações Florestais. **Boletim SNIF 2016**, v.2, ed.2. 2016. Disponível em: http://www.florestal.gov.br/documentos/publicacoes/2230-boletim-snif-producao-florestal-2016/file. Acesso em: 12 jan. 2020.