Universidade de Brasília
Departamento de Estatística

Isabela Paranhos Pinto

Análise e modelagem de dados de afastamento do trabalho por problemas de saúde de servidores públicos federais.

Brasília
2018

Isabela Paranhos Pinto

Análise e modelagem de dados de afastamento do trabalho por problemas de saúde de servidores públicos federais.

Trabalho de Conclusão de Curso
apresentado ao Departamento de Estatística
da Universidade de Brasília, como parte dos
requisitos para a obtenção do título de
Bacharel em Estatística.

Orientador:
Prof. Dr. **Helton Saulo**

Brasília
2018

Resumo

Esse trabalho apresenta uma análise exploratória dos dados de afastamento do trabalho por problemas de saúde dos servidores de um órgão público federal no período de 2006 a 2015. Essa análise consiste em analisar o número de atestados diários e o número de dias que o servidor se encontra ausente de seu trabalho com outras variáveis (Classificação Internacional de Doenças (CID), sexo, idade, departamento de lotação, tipo de vínculo e o período). Além disso, será utilizado Modelos Lineares Generalizados para séries temporais de contagem para modelar a série do número de atestados diários. Será demonstrado o modelo, juntamente com o pacote **tscount** do software R, que fornece fornece métodos de estimativa baseados em verossimilhança para a análise e modelagem de séries temporais de contagem seguindo modelos lineares generalizados. O pacote inclui métodos para o ajuste do modelo e análise de resíduos, previsão e análise de intervenção. Este último será aplicada para detectar uma mudança de nível, na série do número de atestados diários, após a implementação do ponto eletrônico.

PALAVRAS CHAVE: Modelos Lineares Generalizados. Série Temporal para Dados de Contagem. Absenteísmo. Análise Estatística. Setor Público. Teste de Intervenção. Ponto Eletrônico. R. CID.

Abstract

This dissertation presents an exploratory data analysis of the employees moved away from work to a public hospital due to health conditions between 2006 and 2015. This analysis aims to analyze the number of daily health attestations and the number of days that the employee is away from work, related to other variables such as (International Diseases Classification (ISC), gender, age, overload department, types of bonds and the time period). Besides that, Generalized Linear Models will be used for counting temporal series to shape the number series from daily health attestations. The model will be demonstrated along with the package **tscount** of the R software, which provides estimate methods based on verisimilitude to the shaping and analysis of temporal series of counting, obeying Generalized Linear Models. The package includes methods for the adjustments of the model and analysis of residue, prediction and analysis of interventions. The last analysis will be applied to detect a change in level of the number series of daily health attestations, after the implementation of the electronic clock in.

KEYWORDS: Generalized Linear Models. Count Time Series. Absenteeism. Statistical Analysis. Public Sector. Intervention Test. Electronic Clock. R. ISC.

Contents

1 Introdução

O termo absenteísmo é usado para designar as ausências dos trabalhadores no processo de trabalho, seja por falta ou atraso, devido a algum motivo interveniente, sendo assim, é considerado o tempo que o empregador paga pelas atividades produtivas, mas que o empregado não se encontra na empresa.

O absenteísmo pode ser atribuído por causas conhecidas ou causas ignoradas. Dentre as conhecidas estão os casos de férias, casamentos, nascimentos, óbitos e mudanças de domicílio, que são amparadas por lei e assim são justificadas ao empregador, solicitando a permissão de ausência. As ignoradas são justificadas geralmente por problemas de saúde, como: doenças respiratórias, lesões ortopédicas, doenças no aparelho digestivo ou motivos psiquiátricos.

O absenteísmo no setor público é uma realidade como em instituições privadas, no entanto, seu impacto econômico é bastante preocupante, uma vez que gera gastos públicos, afetando toda a população. As organizações públicas apresentam maior quantidade de dias perdidos por absenteísmo-doença, bem como um período maior de afastamento do que as empresas privadas. O servidor estatutário possui uma imagem cercada de mitos a respeito de regalias e isso tem dificultado a compreensão do adoecimento nessas pessoas e a implementação de políticas de saúde efetivas.

A ausência ao trabalho por motivos de saúde é um tema de interesse para o emprego público, e também para os profissionais da saúde. É importante caracterizar o perfil das licenças médicas, assim como identificar os fatores que influenciam os afastamentos dos servidores de um órgão público federal.

É possível observar séries temporais de contagem em várias áreas, aparecem naturalmente sempre que um número de eventos por período de tempo é observado ao longo do tempo. Há inúmeras aplicações no dia-a-dia, como por exemplo, o número diários de internações hospitalares, para a saúde pública, o número de transações na bolsa de valores por minuto, para a economia, ou o número de peças defeituosas por hora, para o controle de qualidade industrial. Modelos para séries temporais de contagem devem levar em consideração que as observações são números inteiros não-negativos e que devem capturar satisfatóriamente a dependência entre as observações. Uma forma conveniente e flexível de abordagem é o emprego da metodologia do Modelo Linear Generalizado (MLG) (Nelder

& Wedderburn, 1972) para modelar condicionalmente as observações com as informações passadas. Essa metodologia é implementada ao escolher uma distribuição apropriada para os dados de contagem e uma função de ligação adequada. Essa abordagem é usada por Fahrmeir (2001) e Kedem & Fokianos (2002), entre outros.

Nesse contexto, esse trabalho tem dois objetivos principais: 1) fazer uma análise exploratória dos dados de afastamento do trabalho por problemas de saúde dos servidores de um órgão público federal no período de 2006 a 2015. Em particular, do número de atestados diários e do número de dias de afastamento, abordando diversas variáveis associadas aos dados (Classificação Internacional de Doenças (CID), sexo, idade, departamento de lotação, tipo de vínculo e o período). Essa análise é importante, para o órgão público federal, porque pode servir de fundamento para a criação de políticas de prevenção com o objetivo de diminuir os afastamentos por motivo de saúde. 2) utilizar MLG para séries temporais de contagem para a modelagem da série do número de atestados diários. Um ponto interessante nessa série é verificar se houve alguma mudança de nível com a implementação do ponto eletrônico, e para isso uma análise de intervenção será aplicada. As análises serão feitas com ajuda do pacote **tscount** (Liboschik *et al.* , 2017) do software livre e de código aberto R (R Core Team, 2016).

Além deste capítulo introdutório, o restante desse trabalho é organizado na seguinte forma; o Capíulo 2 é dedicado a apresentar o banco de dados e uma análise exploratória; no Capítulo 3 é descrito o modelo de MLG para séries temporais de contagem; o Capítulo 4 exibe a análise dos dados do número de atestados diários, utilizando o modelo do Capítulo 3; e, o Capítulo 5 retrata a conclusão do trabalho.

2 Banco de Dados e Análise Exploratória

Esse capítulo tem por objetivo fazer uma análise exploratória do dados da área de saúde dos servidores de um órgão público federal, a fim extrair informações relevantes para a perícia médica deste órgão ou criar políticas de prevenção para melhorar a qualidade no trabalho dos servidores. Além disso, esse capítulo também por objetivo descrever algumas características dos dados e estabelecer relações entre as variáveis do estudo; o número de atestados diários e o número de dias de afastamento. Essa análise exploratória visa fornecer informações sobre o absenteísmo e orientar a formulação de hipóteses. Todas as análises serão realizadas através do software R (R Core Team, 2016).

2.1 Banco de Dados

Os dados utilizados são do banco de dados do Serviço de Perícia Médica de um órgão público federal, onde são registrados os afastamentos por motivo de saúde. Os atestados dos servidores são entregues na perícia médica e são homologados em um sistema interno usado neste órgão.

O banco de dados utilizado nesse trabalho contém informações do número de atestados diários, ou seja, o número de servidores que apresentaram atestado médico em determinado dia, e o número de dias que o servidor irá ficar afastado de seu trabalho por algum problema de saúde. O período abrangido será de janeiro de 2006 até dezembro de 2015, e o mesmo contém as seguintes variáveis:

- Ponto: identificação do servidor;
- Sexo: feminino ou masculino;
- Nascimento: data de nascimento;
- Departamento: departamento de lotação do servidor;
- CID: código da doença;
- Início: dia em que começou o atestado;
- Dia: número de dias do atestado.

2.2 Análise Descritiva

Uma análise das estatísticas descritivas do banco de dados é fundamental para resumirmos algumas informações sobre os mesmos. Iremos utilizar medidas de posições como média, mediana e moda, medidas de dispersão e os quartis. Os quartis são valores dados a partir do conjunto de obervações ordenado em ordem crescente, que dividem a distribuição em quatro partes iguais: o primeiro quartil, Q1, é o número que deixa 25% das observações abaixo e 75% acima, enquanto que o terceiro quartil, Q3, deixa 75% das observações abaixo e 25% acima. Já Q2 é a mediana, deixa 50% das observações abaixo e 50% das observações acima.

Iremos utilizar também o coeficiente de variação (CV%), que é usado para analisar a dispersão em termos relativos a seu valor médio, e ele será dado em porcentagem. Quanto menor for o valor do coeficiente de variação, mais homogêneos serão os dados, ou seja, menor será a dispersão em torno da média. O coeficiente de curtose (CK) caracteriza o "achatamento"da curva da função de distribuição. Se o valor for positivo, dizemos que a função de distribuição é leptocúrtica e possui a curva de distribuição mais afunilada com um pico mais alto do que a distribuição normal, nesse caso dizemos que essa distribuição possui caudas pesadas. E o coeficiente de assimetria (CS), que permite distinguir as distribuições assimétricas: um valor positivo indica que a cauda do lado direito é maior que a do lado esquerdo.

2.2.1 Análise Descritiva do Número de Atestados Diários

Foram observados 29721 afastamentos dos servidores efetivos de um órgão público federal durante o período de 10 anos, perfazendo uma média de 2972 afastamentos por ano. A Tabela 1 mostra o número de atestados, o número de funcionários e a média de atestado por servidor a cada ano, tendo como referência o dia 31 de dezembro. Pode-se observar uma tendência crescente do número de atestados, porém, o número de funcionários se mantém constante, ou seja, ao longo dos anos a ausência no trabalho por motivos de saúde tem aumentado. Pode-se confirmar isso observando o Gráfico 1, que apresenta o número de atestados a cada ano dividido pelo total de atestados.

Tabela 1: Média de Atestados dos Servidores Efetivos

Ano	Número de Atestados	Número de Funcionários	Média de Atestado
2006	1758	3567	0,493
2007	1809	3526	0,513
2008	1933	3489	0,554
2009	2041	3534	0,578
2010	2185	3491	0,626
2011	2437	3408	0,715
2012	2805	3372	0,832
2013	3136	3397	0,923
2014	3789	3396	1,116
2015	7828	3196	2,449

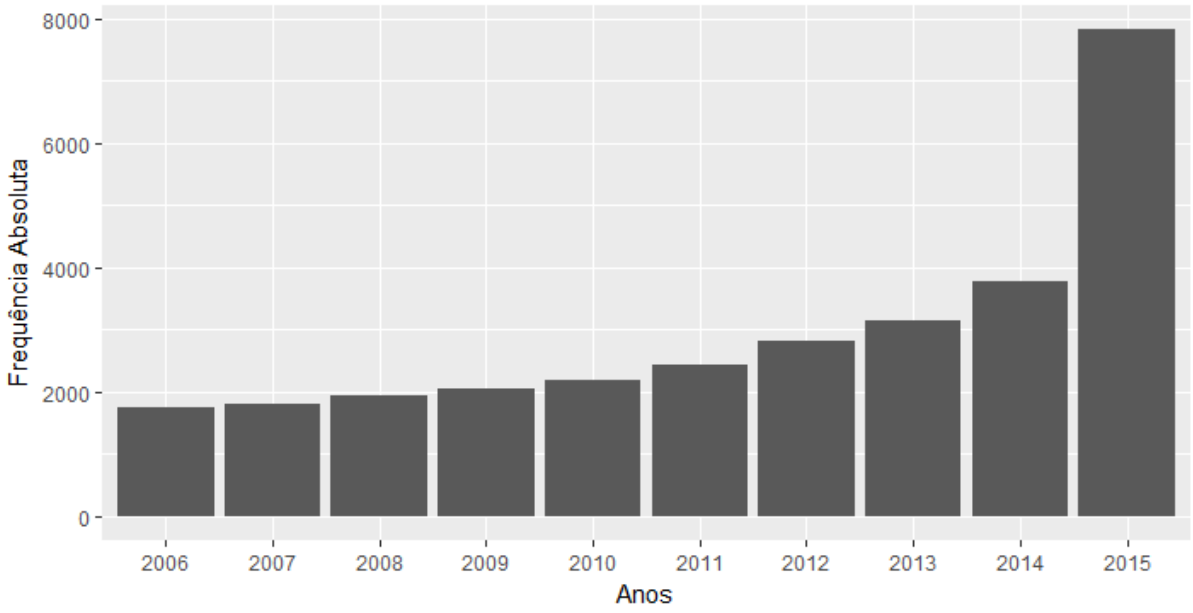


Figura 1: Quantitativo do número de atestados dos servidores públicos efetivos, de 2006 a 2015.

A Figura 2 apresenta o número de atestados para cada ano ao longo dos meses. Pode-se notar que os meses de janeiro, julho e dezembro apresentam um menor número de afastamentos, o que pode ser explicado por ser época de férias, não havendo a necessidade de apresentar o atestado. Observa-se também, que em maio do ano de 2015 houve um aumento no número de afastamentos. Um motivo que poderia explicar esse crescimento significativo é a implementação do ponto eletrônico, que será evidenciado na Seção 2.4.

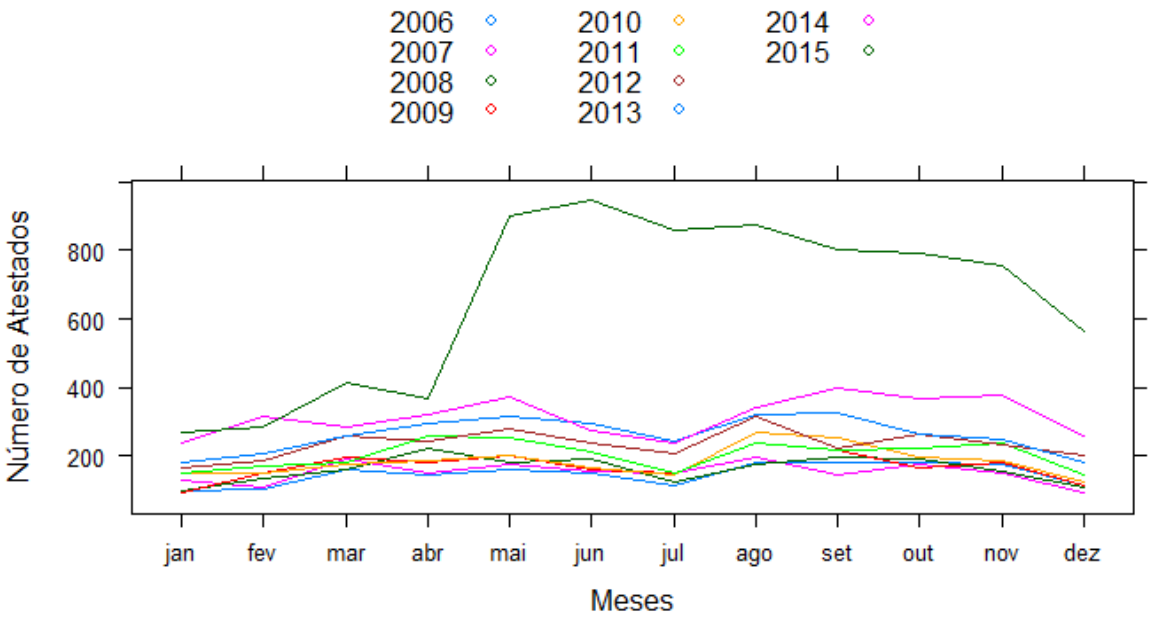


Figura 2: Quantitativo do número de atestados dos servidores públicos efetivos por mês, de 2006 a 2015.

2.2.1.1 CID

A Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde, frequentemente designada pela sigla CID, fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças.

O CID é dividido em capítulos; cada capítulo reúne condições médicas correlacionadas entre si. Nesta análise será utilizada a divisão em capítulos, já que se trata de uma variável qualitativa nominal com muitos níveis.

No Gráfico 3 e na Tabela 2, estão representados os quantitativos de servidores afastados segundo os grupos de doenças que motivaram os afastamentos. Destaca-se como frequente os fatores que influenciam o estado de saúde, que se constituem, predominantemente, das convalescenças pós-cirurgia. Como segundo e terceiro grupo, aparecem doenças do aparelho respiratório e doenças do sistema osteomuscular e do tecido conjuntivo. Lesões, envenenamentos e algumas outras consequências de causas externas ocupam o quarto lugar e os transtornos mentais e comportamentais surge como quinto motivo de afastamento.

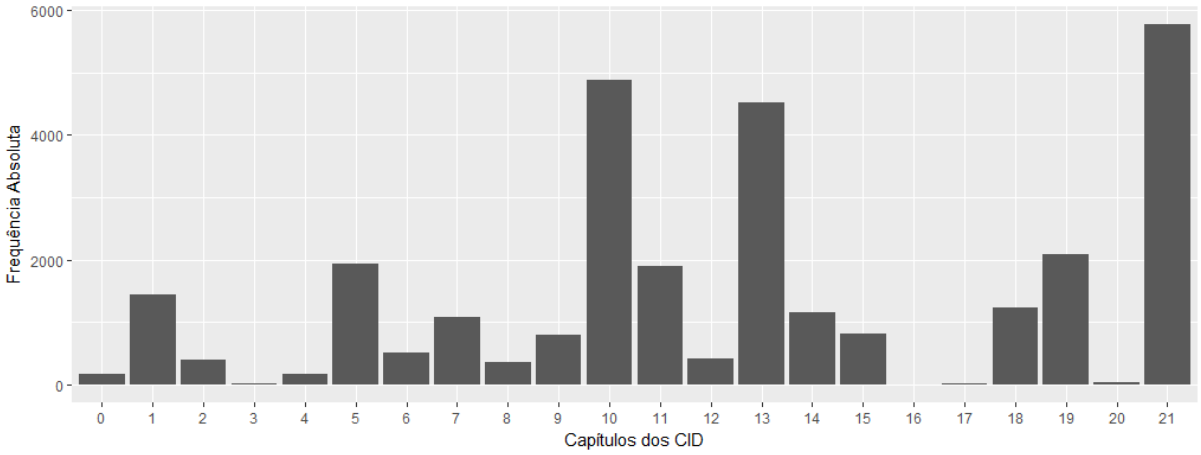


Figura 3: Quantitativo dos servidores públicos efetivos, segundo Capítulos do CID 10, de 2006 a 2015.

Na Tabela 2, estão representados os principais grupos de patologias, em magnitude, que motivaram os afastamentos dos servidores ao longo dos anos. Nesta tabela, pode-se observar a frequência absoluta e a frequência relativa dos 7 grupos mais incidentes, que correspondem a 75.8% das licenças. Enquanto os outros 14 grupos correspondem apenas a 24.2% .

Tabela 2: Frequência absoluta e relativa de acordo com o CID, de 2006 a 2015.

Códigos	Capítulos	Descrição	Freq	Freq (%)
Z00 - Z99	21	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	5770	19.41
J00 - J99	10	Doenças do aparelho respiratório	4870	16.39
M00 - M99	13	Doenças do sistema osteomuscular e do tecido conjuntivo	4522	15.21
S00 - T98	19	Lesões, envenenamentos e algumas outras consequências de causas externas	2086	7.02
F00 - F99	5	Transtornos mentais e comportamentais	1936	6.50
K00 - K93	11	Doenças do aparelho digestivo	1908	6.42
A00 - B99	1	Algumas doenças infecciosas e parasitárias	1442	4.85
		Outros CIDs	7187	24.2

2.2.1.2 Sexo e Idade

Em relação ao gênero da população estudada, constatou-se que do total de 29721 atestados que compõem os servidores efetivos de um órgão público federal, a maioria é do sexo feminino, como mostra a Figura 4. Os dados mostram que 62.1% dos trabalhadores são mulheres, enquanto 37.9% dos servidores são homens.

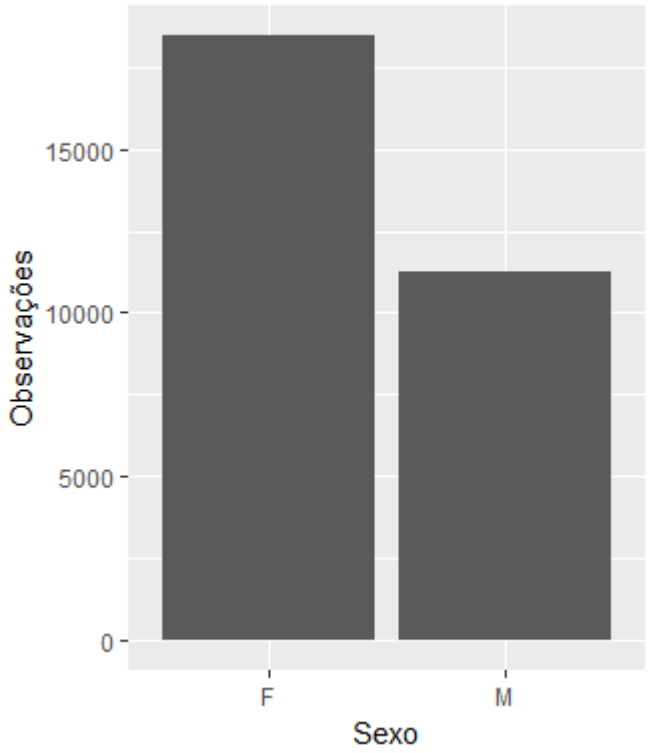


Figura 4: Perfil das licenças dos servidores públicos efetivos, segundo o gênero, de 2006 a 2015.

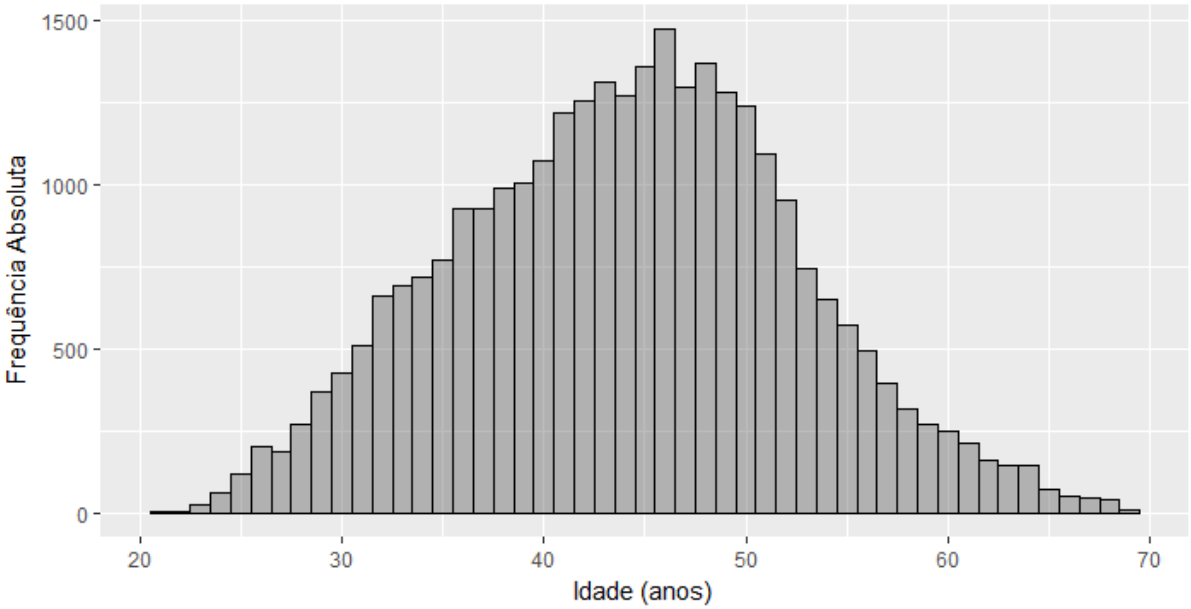


Figura 5: Distribuição da idade dos servidores públicos efetivos que se afastaram do trabalho por motivos de saúde, de 2006 a 2015.

No que diz respeito à idade, a maioria dos servidores tem entre 40 e 50 anos. Falando em termos de quartis, o primeiro quartil, Q1, é de 38 anos e o terceiro quartil, Q3, de 50

anos. A mediana, ou segundo quartil, Q2, é de 44 anos e a média é de 44.1 anos, como a média e a mediana são muito próximas, é considerada que é uma distribuição simétrica. A menor observação é de 21 anos, e a maior é 69 anos, e apenas 25% dos servidores tem menos de 38 anos ou mais de 50 anos. Este fato pode ser entendido por estar sendo analisado os servidores efetivos, ou seja, são servidores que têm como vínculo um concurso público, sendo esperada a inexistência de pontos extremos, como idades muito baixas ou muito altas; além disso, pessoas jovens tendem a ser mais saudáveis, e por consequência, não adoecem tanto quanto os trabalhadores acima dos 40 anos. Pode-se observar isso na figura (5).

2.2.1.3 Departamento

Esse órgão público federal é dividido em departamentos. Na Tabela 3 pode-se observar os 12 departamentos que mais apresentaram afastamentos entre 2006 e 2015. Em primeiro lugar encontra-se o departamento médico; uma possível explicação seria a maior exposição a agentes infecciosos ou ainda as fontes de estresse (físico e psíquico) inerentes à organização do trabalho na área da saúde. Segue-se o Departamento de Comissões em segundo lugar, e em terceiro lugar o Departamento de Taquigrafia, Revisão e Redação. Neste último caso, os servidores estão submetidos a tarefas repetitivas e com pouca flexibilidade, geradoras de lesões osteomusculares e doenças psiquiátricas.

Tabela 3: Frequência absoluta e relativa de acordo com os departamentos, de 2006 a 2015.

Departamento	Freq	Freq (%)
Departamento Médico	3755	12.63
Departamento de Comissões	2854	9.60
Departamento de Taquigrafia, Revisão e Redação	2378	8.00
Centro de documentação e Informação	2212	7.44
Departamento de Polícia Legislativa	1744	5.87
Consultoria Legislativa	1441	4.85
Departamento de Pessoal	1357	4.57
Departamento Técnico	1126	3.79
Departamento de Finanças, Orçamento e Contabilidade	1070	3.60
Departamento de Mídias Integradas	1050	3.53
Centro de Informática	847	2.85
Departamento de Apoio Parlamentar	741	2.49
Departamento de Material e Patrimonio	727	2.45
Outros Departamentos	8419	28.33

2.2.2 Análise Descritiva do Número de Dias de Afastamento

Na Tabela 4 pode-se observar as estatísticas do número de dias de afastamento, isto é, quantos dias o servidor ficou afastado do seu cargo por problemas relacionados à saúde. Nota-se que a média é em torno de 7 dias, exceto no ano de 2015, onde observa-se uma média mais baixa de 5 dias. Consta um desvio padrão elevado, chegando no ano de 2008 a 13 dias. A Figura 6 apresenta a média dos dias de afastamento para cada ano ao longo dos meses, nota-se que é aleatória e não apresenta nenhuma tendência. Essa variação pode ser explicada pelo tipo de doença que gerou o afastamento (aguda ou crônica) e pelo tipo de tratamento requerido. Assim, uma doença aguda, ou seja, de curta duração, gera afastamento breve. Uma doença crônica, isto é, de duração prolongada, ou que exija tratamentos longos ou tóxicos, gera licenças de mais extensas. Nota-se isso na amplitude, pois apresenta atestados de 1 dia até 180 dias.

Tabela 4: Estatísticas do número de dias de afastamento

Anos	Média	Med	Moda	SD	CV (%)	CS	CK	Amplitude	Max	Freq	Soma
2006	6.68	3	1	9.134	136.724	3.969	23.501	89	90	1758	11744
2007	8.296	4	2	11.677	140.765	3.643	17.957	89	90	1809	15007
2008	8.443	4	2	13.241	156.819	5.064	39.029	179	180	1933	16321
2009	6.737	3	2	10.817	160.56	7.44	93.015	179	180	2041	13750
2010	7.677	3	2	11.981	156.063	5.315	44.769	178	179	2185	16775
2011	6.414	3	1	9.191	143.279	3.991	25.17	119	120	2437	15632
2012	7.191	3	1	11.873	165.11	5.743	54.079	179	180	2805	20171
2013	7.088	3	1	11.859	167.299	5.928	57.988	179	180	3136	22229
2014	6.932	3	1	9.798	153.273	5.449	55.503	179	180	3789	24220
2015	5.229	2	1	9.701	185.503	6.772	78.913	179	180	7828	40935
Geral	6.621	3	1	10.785	162.895	5.715	55.967	179	180	29721	196784

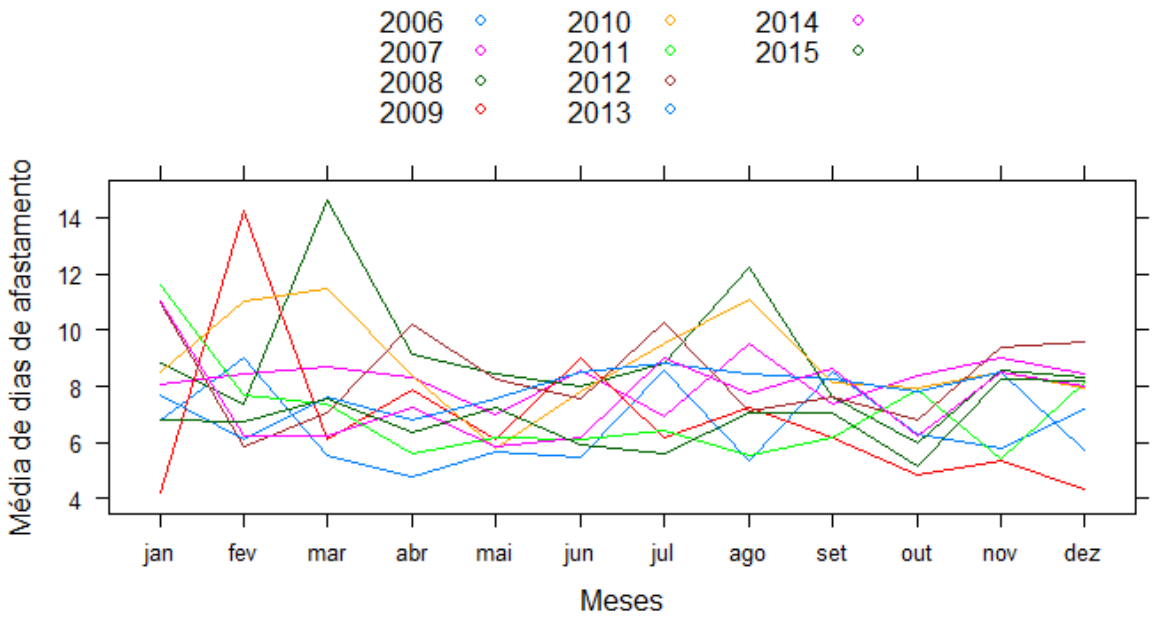


Figura 6: Quantitativo da média de dias de afastamento dos servidores públicos efetivos por mês, de 2006 a 2015.

2.2.2.1 CID

A Tabela 5 apresenta, em ordem decrescente , os capítulos dos CIDs, que em média de dias, afastaram por mais tempo os servidores de seus serviços. Nota-se que algumas doenças geram afastamentos longos, apesar de serem doenças com alta frequência de licenças. É possível entender este fato, observando, como exemplo, o capítulo que compreende as neoplasias malignas, que geram maior debilidade e exigem tratamentos prolongados. Em segundo lugar, tem-se o capítulo 19; sua alta frequência e duração podem ser explicadas pelas fraturas e torções, que estão incluídas nesse grupo. Em terceiro lugar, encontra-se o CID F, que representa, na grande maioria, as doenças psiquiátricas.

Tabela 5: Média de dias de afastamento, por CID, de 2006 a 2015.

Códigos	Capítulos	Descrição	Média de Dias	Freq.
C00-D48	2	Neoplasmas (tumores)	20.3	392
S00-T98	19	Lesões, envenenamentos e algumas consequência de causas externas	12.3	2086
F00-F99	5	Transtornos mentais e comportamentais	12.0	1936
I00-I99	9	Doenças do aparelho circulatório	9.23	804
O00-O99	15	Gravidez, parto e puerpério	9.20	813
E00-E90	4	Doenças endócrinas, nutricionais e metabólicas	8.88	170
G00-G99	6	Doenças do sistema nervoso	8.70	505
M00-M99	13	Doenças do sistema osteomuscular e do tecido conjuntivo	8.24	4522

2.2.2.2 Sexo e Idade

As Tabelas 6 e 7 apresentam as estatísticas descritivas do número de dias de afastamento para cada sexo. Pode-se observar que o sexo masculino tem atestados mais longos que o sexo feminino, e um desvio padrão maior também.

Tabela 6: Estatísticas do número de dias para o Sexo Feminino

Anos	Média	Med	Moda	SD	CV (%)	CS	CK	Amplitude	Max	Freq
2006	6.61	3	1	8.816	133.36	4.22	27.837	89	90	1101
2007	8.269	4	2	11.293	136.563	3.519	17.307	89	90	1121
2008	7.814	3	2	11.211	143.459	4.769	39.196	156	157	1159
2009	6.635	3	2	9.998	150.676	6.691	84.201	179	180	1239
2010	7.089	3	1	10.738	151.479	6.397	73.91	178	179	1336
2011	5.982	3	1	7.906	132.155	3.046	12.092	68	69	1520
2012	6.772	3	2	10.544	155.712	5.851	61.496	179	180	1721
2013	7.083	3	1	10.398	146.803	4.086	25.514	119	120	1972
2014	5.858	3	1	8.283	141.398	4.033	27.859	119	120	2399
2015	4.964	2	1	8.327	167.745	4.949	39.097	136	137	4895
Geral	6.303	3	1	9.521	151.073	4.977	45.539	179	180	18463

Tabela 7: Estatísticas do número de dias para o Sexo Masculino

Anos	Média	Med	Moda	SD	CV (%)	CS	CK	Amplitude	Max	Freq
2006	6.798	3	1	9.649	141.946	3.608	17.907	87	88	657
2007	8.339	4	2	12.286	147.344	3.771	18.334	89	90	688
2008	9.385	4	2	15.762	167.945	4.839	32.328	179	180	774
2009	6.894	3	2	11.978	173.738	7.931	93.923	179	180	802
2010	8.603	4	2	13.667	158.86	4.257	23.141	119	120	849
2011	7.131	3	2	10.962	153.721	4.263	26.377	119	120	917
2012	7.857	3	1	13.7	174.369	5.38	43.845	179	180	1084
2013	7.097	3	2	13.994	197.181	6.878	66.99	179	180	1164
2014	7.314	3	1	11.916	162.913	5.799	56.026	179	180	1390
2015	5.672	2	1	11.625	204.962	7.411	83.627	179	180	2933
Geral	7.143	3	1	12.569	175.958	5.931	54.54	179	180	11258

2.2.2.3 Departamento

A Tabela 8 apresenta as estatísticas descritivas do número de dias de afastamento para os departamentos que mais apresentaram afastamentos entre 2006 e 2015.

Tabela 8: Estatística do número de dias por Departamento (os 70% mais frequentes)

Departamento	Média	Med	Moda	SD	CV (%)	CS	CK	Amplitude	Max	Freq
Departamento Médico	4.912	2	1	8.53	173.666	6.005	69.169	179	180	3755
Departamento de Comissões	6.97	3	1	11.297	162.08	5.632	54.101	179	180	2854
Departamento de Taquigrafia, Revisão e Redação	6.05	3	1	9.755	161.251	5.463	45.81	136	137	2378
Centro de documentação e Informação	5.779	3	1	8.19	141.72	5.362	51.807	136	137	2212
Departamento de Polícia Legislativa	8.233	3	1	14.501	176.127	4.608	28.947	179	180	1744
Consultoria Legislativa	7.543	3	1	11.082	146.91	3.705	19.693	112	113	1441
Departamento de Pessoal	6.966	3	1	11.964	171.751	7.914	99.216	179	180	1357
Departamento Técnico	7.351	3	1	12.983	176.618	6.633	66.719	179	180	1126
Departamento de Finanças, Orçamento e Contabilidade	6.256	3	1	8.572	137.014	2.987	11.099	56	60	1070
Departamento de Mídias Integradas	6.32	3	1	9.646	152.634	4.986	36.319	111	112	1050
Centro de Informática	5.824	3	1	9.993	171.586	5.603	44.292	119	120	847
Departamento de Apoio Parlamentar	7.144	3	1	11.581	162.098	6.418	71.835	178	179	741

2.3 CNE e SP

Esse órgão público federal possui dois tipos de vínculos. Têm os servidores ocupantes de cargos efetivos, que são os concursados e os servidores de cargos em comissão, que são os nomeados por autoridade competente, sem a necessidade de concurso público.

Dentre os servidores de cargos em comissão, tem o secretariado parlamentar (SP), cuja a função é prestar serviços de secretaria, assistência e assessoramento direto e exclusivo nos gabinetes dos deputados. E, os ocupantes de cargos de natureza especial (CNEs) que têm por finalidade a prestação de serviços de assessoramento exclusivamente à Mesa e às Suplências, às Lideranças, às Comissões, à Procuradoria Parlamentar, à Ouvidoria Parlamentar, ao Conselho de Ética e Decoro Parlamentar, à Liderança da Minoria no Congresso, à Procuradoria Especial da Mulher e aos órgãos administrativos da Casa.

Em 2014, esse órgão público federal possuía 11.472 servidores ocupantes em cargos comissionados, 3396 servidores efetivos e foram apresentados 772 e 3789 atestados por motivos de saúde, respetivamente. No ano de 2015, tinham 12381 servidores comissionados, 3196 efetivos e foram apresentados 1232 e 7828 atestados, respectivamente.

Na Tabela 9 pode-se observar que a proporção de afastamentos por servidor dos efetivos é maior que as dos comissionados, em 2015 chega a ser 24 vezes maior.

Tabela 9: Média de afastamentos de acordo com o vínculo e ano

	Vínculo	
Ano	Efetivos	Comissionados
2014	1.116	0.067
2015	2.449	0.100
Total	1.762	0.084

2.4 Ponto Eletrônico

Em maio de 2015 foi implementado o ponto eletrônico nesse órgão público federal. O registro de presença é feito por leitores biométricos que utilizam impressão digital. Os servidores podem optar por uma jornada diária de 7 horas ininterruptas, no total de 35 horas semanais, ou por 8 horas diárias, com intervalo para almoço, no total de 40 horas semanais. Esse registro eletrônico de presença e a regra para as horas extras atingem os servidores efetivos, os cargos de natureza especial, os chamados CNEs, e os secretários parlamentares que atuam em Brasília. Anteriormente a esta data, o controle de frequência era feito mediante registro manual em livros de frequência e em alguns departamento era inexistente.

Tabela 10: Média de atestados por dia, de acordo com o vínculo e o ponto eletrônico

	Vínculo	
Ponto Eletrônico	Efetivos	Comissionados
Antes	6.817	0.285
Depois	26.514	4.216

Pode-se notar na Tabela 10 um aumento significativo de afastamentos após a implantação do ponto eletrônico. Antes do registro eletrônico, no período de janeiro de 2006 até abril de 2015, a média de atestados para os servidores efetivos era de 6.8 atestados por servidor e esse número quase quadruplicou após o ponto eletrônico, o período que corresponde a maio de 2015 à dezembro de 2015, 245 dias. Observa-se um aumento ainda maior nos servidores comissionados.

Um dos motivos que poderia explicar este aumento significativo, seria a informalidade gerada pelo registro manual em caso de necessidade de afastamento por motivo de saúde.

Antes da implementação do ponto eletrônico, os servidores possuíam maior flexibilidade no cumprimento de sua jornada de trabalho e controle de seus horários. Com o registro eletrônico estabeleceu-se uma forma rígida de controle de horário, o que pode gerar a perda de autonomia dos trabalhadores. É fato conhecido na literatura médica que a autonomia no trabalho é associada ao bem estar e, por outro lado, a rigidez e falta de controle sobre o processo de trabalho associam-se a estresse e adoecimento.

É importante a análise estatística desses dados, de forma a confirmar, ou não, as diferenças observadas na análise descritiva.

3 MLG para séries temporais de contagem

Este Capítulo tem por finalidade explicar o modelo de MLG para séries temporais de contagem proposto por Liboschik *et al.* (2017). Na Seção 3.1 é apresentado os modelos que considerados, de MLG e de séries temporais; a Seção 3.2 descreve a estimativa de quase máxima verossimilhança (EMV) dos parâmetros do modelo desconhecido; a Seção 3.3 é sobre a previsão com esses modelos; a Seção 3.4 resume as ferramentas para a validação do modelo; e, a Seção 3.5 discute procedimentos para detectar intervenções.

3.1 Modelo

Seja $\{Y_t : t \in \mathbb{N}\}$ uma série temporal de contagem. Vamos definir $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,r})^\top$ sendo um vetor de covariáveis r-dimensional variável no tempo, e $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = \lambda_t$ sendo a média condicional da série de contagem, na qual $\lambda_t : t \in \mathbb{N}$. Simbolizamos por $\mathcal{F}_t$ o processo conjunto $\{Y_t, \lambda_t, \mathbf{X}_{t+1} : t \in \mathbb{N}\}$ até o tempo t incluindo a informação de covariáveis no tempo $t+1$. O pressuposto da distribuição para Y_t dado $\mathcal{F}_{t-1}$ será discutido depois. Agora, estamos interessados na forma geral do modelo

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l \, g(\lambda_{t-j_l}) + \boldsymbol{\eta}^\top \mathbf{X}_t, \tag{1}$$

em que $g : \mathbb{R}^+ \to \mathbb{R}$ é uma função de ligação e $\tilde{g} : \mathbb{N}_0 \to \mathbb{R}$ é uma função de transformação. O vetor de parâmetros $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)^\top$ corresponde ao efeito das covariáveis. Denominamos $v_t = g(\lambda_t)$ de preditor linear, na nomenclatura de MLGs. Para permitir a regressão em observações passadas de respostas arbitrárias, definimos um conjunto $P = \{i_1, i_2, \ldots, i_p\}$, no qual $0 < i_1 < i_2 \ldots < i_p < \infty$ e inteiros, e $p \in \mathbb{N}_0$. Isso nos permite regredir na observações passadas $Y_{t-i_1} < Y_{t-i_2} \ldots < Y_{t-i_p}$. Analogamente, definimos um conjunto $Q = j_1, j_2, \ldots, j_q$, $q \in \mathbb{N}_0$ e $0 < j_1 < j_2 \ldots < j_q < \infty$ inteiros para regressão de médias condicionais defasadas $\lambda_{t-j_1} < \lambda_{t-j_2} \ldots < \lambda_{t-j_q}$. Através da teoria dos modelos temos, podemos escolher os melhores valores para p e q definindo alguns parâmetros do modelo para zero, em que $P = \{1, \ldots, p\}$ e $Q = \{1, \ldots, q\}$. Definimos os conjuntos P e Q, ou seja, especificamos a ordem do modelo considerando as funções empíricas de autocorrelação dos dados observados.

Considere a situação em que g e $\tilde{g}$ são iguais a identidade, $P = 1, \ldots, p$, $Q = 1, \ldots, q$

e $\boldsymbol{\eta} = 0$. Assim, o Modelo 1 fica

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{l=1}^{q} \alpha_l \lambda_{t-l}. \tag{2}$$

Assumindo que Y_t dado o passado segue uma distribuição de Poisson, obtemos um modelo GARCH de valor inteiro de ordem p e q, que pode ser abreviado por INGARCH(p, q). Esse modelo também é conhecido por modelos condicionais autoregressivos de Poisson (ACP). Ele foram discutidos por Heinen (2003); Ferland *et al.* (2006); Fokianos *et al.* (2009), entre outros. Quando $\boldsymbol{\eta} \neq 0$ temos um modelo INGARCH com covariáveis não-negativas, e é feito dessa forma porque precisamos garantir que o processo médio resultante seja positivo.

Vamos considerar novamente o Modelo 1, mas agora a função de ligação será uma função logarítmica, $g(x) = log(x), \tilde{g}(x) = log(x + 1)$ e P, Q como antes. Agora temos um modelo log-linear para análise de série temporal de contagem de ordem p e q. Definindo $v_t = log(\lambda_t)$, temos:

$$v_t = \beta_0 + \sum_{k=1}^{p} \beta_k log(Y_{t-k} + 1) + \sum_{l=1}^{q} \alpha_l v_{t-l}. \tag{3}$$

Este modelo log-linear é estudado por Fokianos & Tjøstheim (2011); Woodard *et al.* (2011) e Douc *et al.* (2013). Liboschik *et al.* (2017) seguiu Fokianos & Tjøstheim (2011) utilizando a função $\tilde{g}(x) = log(x+1)$ na transformação de observações passadas, assim ela e o preditor linear v_t ficam na mesma escala. Sem afetar na inferência, pode-se adicionar uma constante c em cada observação para evitar valores iguais a zeros; uma escolha razoável para c seria 1. Observe que no Modelo 3 pode ser modelada uma correlação serial negativa, enquanto no Modelo 2 apenas a positiva. Além disso, o Modelo 3 acomoda covariáveis mais facilmente do que o Modelo 2, uma vez que o modelo log-linear implica positividade do processo de média condicional λ_t. O Modelo Linear 2 com as covariáveis deve receber um cuidado especial, pois está limitado a efeitos positivos em λ_t. No Modelo 3 os efeitos das covariáveis na resposta são multiplicativos, enquanto no Modelo 2 é aditivo.

No Modelo 1 juntamente com a suposição de Poisson, ou seja, $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$, implica que

$$P(Y_t = y|\mathcal{F}_{t-1}) = \frac{\lambda_t^y exp(-\lambda_t)}{y!}, \qquad y = 0, 1, \ldots \tag{4}$$

Esse modelo tem $\text{VAR}(Y_t|\mathcal{F}_{t-1}) = \mathbb{E}(Y_t|\mathcal{F}_{t-1}) = \lambda_t$. Logo, no caso do modelo de resposta condicional de Poisson, a média condicional é igual a variância do processo observado.

Na distribuição da Binomial Negativa a variância é maior que a média λ_t, o que é chamado de superdispersão. Segundo Christou & Fokianos (2014), assumi-se que $Y_t|\mathcal{F}_{t-1} \sim \text{BN}(\lambda_t, \phi)$, em que $\phi \in (0, \infty)$ e é o parâmetro adicional de dispersão. Então,

$$P(Y_t = y|\mathcal{F}_{t-1}) = \frac{\Gamma(\phi + y)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_t}\right)^{\phi} \left(\frac{\lambda_t}{\phi + \lambda_t}\right)^{y}, \qquad y = 0, 1, \ldots \cdot \tag{5}$$

Nesse caso, $\text{VAR}(Y_t|\mathcal{F}_{t-1}) = \lambda_t + \lambda_t^2/\phi$, isto é, a cada λ_t a variância condicional aumenta quadraticamente. A distribuição de Poisson é um caso limitante da Binomial Negativa quando $\phi \to \infty$. Observe que a distribuição Binomial Negativa pertence à classe de processos mistos de Poisson. Um processo misto de Poisson é definido por $Y_t = N_t(0, Z_t\lambda_t]$, no qual $\{N_t\}$ é i.i.d. e segue uma Poisson com intensidade unitária e $\{Z_t\}$ é uma variável aleatória i.i.d. com média 1 e variância σ^2, independente de $\{Y_t\}$ (Christou & Fokianos (2014)). Quando $\{Z_t\}$ é um processo de varáveis aleatórias i.i.d. que seguem uma Gamma, obtemos uma Binomial Negativa com $\sigma^2 = 1/\phi$. Liboschik *et al.* (2017) referem-se a σ^2 como o coeficiente de superdispersão porque é proporcional à extensão da superdispersão da distribuição condicional. O caso limite de $\sigma^2 = 0$ corresponde à distribuição de Poisson, ou seja, sem superdispersão. O procedimento de estimativa que foi estudado não está confinado ao caso Binomial Negativo, mas a qualquer distribuição de Poisson mista. Entretanto, a suposição de ser Binomial Negativa é necessária para intervalos de predição e análise de resíduos; esses tópicos são discutidos nas Seções 3.3 e 3.4.

No Modelo 1 o efeito de uma covariável entra completamente na dinâmica do processo e se propaga para futuras observações tanto pela regressão em observações passadas quanto pela regressão em médias condicionais passadas. O efeito de tais covariáveis pode ser visto como uma influência interna no processo de geração de dados, e é por isso que Liboschik *et al.* (2017) referem-se a ele como um efeito interno de covariável. Também permite-se incluir covariáveis de modo que seu efeito só se propague a observações futuras pela regressão em observações passadas, mas não diretamente pela regressão das médias condicionais passadas. Seguindo Liboschik *et al.* (2016), para o caso dos efeitos de inter-

venção descritos por covariáveis determinísticas, nos referimos ao efeito de tais covariáveis como um efeito de covariável externo. Seja $\mathbf{e} = (e_1, \ldots, e_r)^\top$ um vetor especificado pelo usuário com $e_i = 1$ se o i-ésimo componente do vetor da covariável tiver um efeito externo e $e_i = 0$ caso contrário, $i = 1, \ldots, r$. Denote por diag($\mathbf{e}$) uma matriz diagonal com elementos diagonais dada por $\mathbf{e}$. Segundo Liboschik *et al.* (2017) a generalização do Modelo 1 permite efeitos internos e externos da covariável, e é dada por

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l (g(\lambda_{t-j_l}) - \boldsymbol{\eta}^T \, diag(\mathbf{e}) \, \mathbf{X}_{t-j_l}) + \boldsymbol{\eta}^T \mathbf{X}_t \cdot \qquad (6)$$

Basicamente, o efeito de todas as covariáveis com efeito externo é subtraído nos termos de retorno, de modo que seu efeito entra na dinâmica do processo apenas por meio das observações.

3.2 Estimação e inferência

O pacote **tscount** fornece métodos de estimação pela máxima verossimilhança (MV) para a análise e modelagem de séries temporais de dados de contagem seguindo modelos lineares generalizados, ou seja, modelos na forma do Modelo 1 (function `tsglm`). Se a suposição de Poisson for verdadeira, então obtemos um estimador comum de MV. No entanto, sob a hipótese de Poisson mista, obtemos um estimador quase-MV. Seja $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r)^\top$ o vetor dos parâmetros de regressão. Independentemente do pressuposto da distribuição, o espaço paramétrico para o Modelo 2 (INGARCH) com covariáveis é dado por

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : \beta_0 > 0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1 \right\}.$$

O intercepto β_0 deve ser positivo e todos os outros parâmetros devem ser não-negativos para garantir a positividade da média condicional λ_t. O espaço paramétrico para o Modelo 3 (log-linear) com covariáveis, é dado por

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : |\beta_1|, \ldots, |\beta_p|, |\alpha_1|, \ldots, |\alpha_q| < 1, \left| \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l \right| < 1 \right\}.$$

Christou & Fokianos (2014) ressaltou que, com a parametrização 5 da distribuição Binomial Negativa, a estimação dos parâmetros de regressão $\boldsymbol{\theta}$ não depende do parâmetro de dispersão adicional ϕ. Isto permite empregar uma abordagem de quase máxima verossimilhança com base na probabilidade de Poisson para estimar os parâmetros de regressão $\boldsymbol{\theta}$, e o parâmetro da dispersão ϕ é estimado separadamente. Essa abordagem é diferente de uma estimativa de máxima verossimilhança baseada na distribuição Binomial Negativa, que por exemplo foi implementada na função `glm.nb` no pacote R **MASS** (Venables & Ripley 2002). Nessa abordagem, a estimação através do método de máxima verossimilhança de uma Binomial Negativa para um parâmetro de dispersão estimado ϕ e a estimativa de ϕ dado os parâmetros de regressão estimados $\boldsymbol{\theta}$ são iterados até a convergência. A abordagem binomial quase-negativa foi escolhida pela simplicidade e sua utilidade em derivar estimadores consistentes quando o modelo para λ_t foi corretamente especificado (Ahmad & Francq 2016).

O vetor de escores da função de log-verossimilhança e a matriz de informação são derivados condicionalmente dos valores pré-amostrais da série temporal e do processo de médias condicionais λ_t, precisamente em $\mathcal{F}_0$. Para um vetor de observações $\mathbf{y} = (y_1, \ldots, y_n)^\top$, a função de quase verossimilhança condicional é dada por

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_t(y_t; \boldsymbol{\theta}) = \sum_{t=1}^{n} \left(y_t \ln(\lambda_t(\boldsymbol{\theta})) - \lambda_t(\boldsymbol{\theta}) \right), \tag{7}$$

em que $p_t(y; \boldsymbol{\theta}) = P(Y_t = y | \mathcal{F}_{t-1})$ é a função de densidade de probabilidade da distribuição de Poisson, como foi definida em 4. A média condicional é considerada uma função $\lambda_t : \Theta \to \mathbb{R}^+$, e assim é denotado por todo $\lambda_t(\boldsymbol{\theta})$ para todo t. A função de escore condicional é o vetor $(p + q + r + 1)$-dimensional que é dado por

$$S_n(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}. \tag{8}$$

O vetor das derivadas parciais $\partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ é calculado recursivamente por recursões, e assim a matriz de informação condicional é dada por

$$G_n(\boldsymbol{\theta};\sigma^2) = \sum_{t=1}^{n} \text{COV}\left(\frac{\partial l(\boldsymbol{\theta};Y_t)}{\partial \boldsymbol{\theta}}\middle|\mathcal{F}_{t-1}\right) = \sum_{t=1}^{n}\left(\frac{1}{\lambda_t(\boldsymbol{\theta})}+\sigma^2\right)\left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\top}.$$

No caso da Poisson temos $\sigma^2 = 0$ e na Binomial Negativa $\sigma^2 = 1/\phi$. Para facilitar a notação, vamos usar a matriz de informação condicional da Poisson, que é $G_n^*(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta};0)$.

O estimador de máxima verossimilhança (EMV) $\widehat{\boldsymbol{\theta}}_n$ de $\boldsymbol{\theta}$, assumindo que ele exista, é a solução do problema de otimização restrito não linear

$$\widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta})\cdot \tag{9}$$

Simbolizamos os valores ajustados por $\widehat{\lambda}_t = \lambda_t(\widehat{\boldsymbol{\theta}})$. Seguindo Christou & Fokianos (2014), o parâmetro de dispersão ϕ da distribuição da Binomial Negativa é estimado pela seguinte equação

$$\sum_{t=1}^{n}\frac{(Y_t-\widehat{\lambda}_t)^2}{(\widehat{\lambda}_t+\widehat{\lambda}_t^2/\widehat{\phi})} = n-(p+q+r+1), \tag{10}$$

que é baseada na estatística χ^2 de Pearson. O parâmetro da variância é estimado por $\widehat{\sigma}^2 = 1/\widehat{\phi}$. Para a distribuição de Poisson temos $\widehat{\sigma}^2 = 0$.

A inferência para os parâmetros de regressão é baseada na normalidade assintótica do EMV. Para um processo de covariância bem comportado $\mathbf{X}_t$ temos

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right) \xrightarrow{d} N_{p+q+r+1}\left(0, G_n^{-1}(\widehat{\boldsymbol{\theta}}_n;\widehat{\sigma}^2)G_n^*(\widehat{\boldsymbol{\theta}}_n)G_n^{-1}(\widehat{\boldsymbol{\theta}}_n;\widehat{\sigma}^2)\right) \tag{11}$$

com $n\to\infty$, em que θ_0 é o verdadeiro valor do parâmetro e $\widehat{\sigma^2}$ é um estimador consistente para σ^2. A informaç!ao sobre cada covariável cresce linearmente com o tamanho da amostra e as covariáveis não são linearmente dependentes.

Um método alternativo à aproximação da normal 11 para a obtenção do erro padrão e intervalo de confiança (function `se`), incluimos um procedimento de boostrap paramétrico (argument `B`), para o qual o tempo de computação muitas vezes é maior. Dessa forma, B séries temporais são simuladas a partir do modelo ajustado aos dados originais. Os erros padrão empíricos da estimativa dos parâmetros para essas B séries temporais são os erros padrão do bootstrap. Os intervalos de confiança são baseados em quantis da amostra de

bootstrap. Esse procedimento pode calcular erros padrão e intervalos de confiança para $\widehat{\theta}$ e $\widehat{\sigma}^2$. Normalmente, utilizando B = 500 gera resultados estáveis.

3.3 Predição

Segundo Liboschik *et al.* (2017) a esperança condicional λ_{n+1} dada no Modelo 1, é o preditor ideal de $\widehat{Y}_{n+1}$ para Y_{n+1} dado $\mathcal{F}_n$, ou seja, todo o processo até o tempo n e as covariáveis potenciais no tempo $n+1$, isso em termos do erro quadrático médio. (`S3` method of function `predictor`). Pela estrutura do modelo, a distribuição condicional de $\widehat{Y}_{n+1}$ é uma distribuição Poisson, Modelo 4, e com uma distribuição condicional da média λ_{n+1} é uma distribuição Binomial Negativa, Modelo 5. O preditor de $\widehat{Y}_{n+h}$ para Y_{n+h} é obtido através das diversas predições de $\widehat{Y}_{n+1}$, em que os valores não observados $Y_{n+1}, \ldots, Y_{n+h-1}$ são substituídos por suas respectivas previsões de $\widehat{Y}_{n+1}$, $h \in \mathbb{N}$. A distribuição desta previsão $\widehat{Y}_{n+h}$ não é conhecida analiticamente, mas pode ser aproximada numericamente por bootstrap que é descrito abaixo.

Nas aplicações λ_{n+1} é substituído pelo seu estimador $\widehat{\lambda}_{n+1} = \lambda_{n+1}(\widehat{\boldsymbol{\theta}})$, que depende dos parâmetros de regressão estimados por $\boldsymbol{\theta}$. O parâmetro de dispersão ϕ da Binomial Negativa é substituído pelo seu estimador $\widehat{\phi}$. Liboschik *et al.* (2017) observa que ao acrescentar os parâmetros estimados no modelo, induz incerteza adicional à distribuição preditiva. Esta incerteza da estimativa não é levada em conta para a construção dos intervalos de predição descritos nos parágrafos seguintes.

Intervalos de predição para Y_{n+1} (argument `level`) com uma determinada taxa de cobertura $1-\alpha$, são projetados para cobrir a verdadeira observação Y_{n+h} com probabilidade de $1-\alpha$. Intervalos simultâneos de previsão para todos $Y_{n+1}, \ldots, Y_{n+h}$ podem ser obtidos pelo teste de Bonferroni com nível de confiança de $1 - \alpha/h$ para cada um (argument `global = TRUE`).

Liboschik *et al.* (2017) se remete aos dois princípios diferentes para a construção de intervalos de predição disponíveis que, na prática, geralmente produzem intervalos idênticos. Em primeiro lugar, os limites podem ser o quantil $(\alpha/2)$ e $(1-\alpha/2)$ da distribuição preditiva (aproximada) (argument `type = "quantiles"`). Em segundo lugar, os limites podem ser escolhidos de tal forma que o intervalo tenha comprimento mínimo, dado que, de acordo com a distribuição preditiva (aproximada), a probabilidade de um valor cair nesse intervalo é pelo menos tão grande quanto a taxa de cobertura desejada

$1-\alpha$ (argument `type = "shortest"`).

Os intervalos de predição para $\widehat{Y}_{n+1}$ podem ser obtidos diretamente da distribuição condicional (argument `method = "conddistr"`). Intervalos de predição obtidos por um procedimento de bootstrap paramétrico (argument `method = "bootstrap"`) são baseados em B simulações de realizações $y_{n+1}^{(b)}, \ldots, y_{n+h}^{(b)}$ do modelo ajustado, $b = 1, \ldots, B$ (argument B). Para obter um intervalo de predição aproximado para Y_{n+h}, pode-se usar os quantis empíricos $(\alpha/2)$ e $(1-\alpha/2)$ de $y_{n+h}^{(1)}, \ldots, y_{n+h}^{(B)}$ (se `type = "quantiles"`) ou encontrar o intervalo mais curto que contém pelo menos $\lceil (1-\alpha) \cdot B \rceil$ dessas observações (se `type = "shortest"`).

3.4 Análise de Resíduos

Liboschik *et al.* (2017) afirma que ferramentas originalmente desenvolvidas para modelos lineares generalizados, bem como para séries temporais, podem ser utilizadas para avaliar o ajuste do modelo e seu desempenho preditivo. Dentro da classe de contagem de séries temporais seguindo modelos lineares generalizados, é necessário avaliar a especificação do preditor linear, bem como a escolha da função de ligação e da distribuição condicional. As ferramentas apresentadas nesta seção facilitam a seleção de um modelo adequado para um certo conjunto de dados. Todas as ferramentas são introduzidas como versões na amostra, o que significa que as observações $y_1, \ldots, y_n$ são usadas para ajustar o modelo, bem como para avaliar o ajuste obtido.

Liboschik *et al.* (2017) continua afirmando que os valores ajustados são denotados por $\widehat{\lambda}_t = \lambda_t(\widehat{\theta})$. Isso não depende da distribuição escolhida, porque a média é a mesma, independentemente da distribuição da resposta. Existem vários métodos para analisar os resíduos (S3 method of function `residuals`).

Os resíduos da resposta (ou brutos) (argument `type = "response"`) são dados por

$$r_t = y_t - \widehat{\lambda}_t, \tag{12}$$

temos os resíduos de Pearson, que é uma alternativa padronizada (argument `type = "Pearson"`)

$$r_t^P = \frac{(y_t - \widehat{\lambda}_t)}{\sqrt{\widehat{\lambda}_t + \widehat{\lambda}_t^2 \, \widehat{\sigma^2}}}, \tag{13}$$

ou os resíduos padronizados Anscombe, distribuídos de forma mais simétrica (argument `type = "anscombe"`)

$$r_t^A = \frac{3/\widehat{\sigma}^2((1+y_t\widehat{\sigma}^2)^{2/3} - (1+\widehat{\lambda}_t\widehat{\sigma}^2)^{2/3}) + 3(y_t^{2/3} - \widehat{\lambda}_t^{2/3})}{2(\widehat{\lambda}_t + \widehat{\lambda}_t^2\,\widehat{\sigma}^2)^{1/6}}, \tag{14}$$

para $t = 1, \ldots, n$ (olhe o exemplo Hilbe (2011), seção 5.1). A função de autocorrelação empírica desses resíduos é útil para diagnosticar a dependência serial que não foi explicada pelo modelo ajustado. Um gráfico dos resíduos em relação ao tempo pode revelar mudanças no processo de geração de dados ao longo do tempo. Além disso, uma plotagem dos resíduos quadrados r_t^2 e dos valores correspondentes $\widehat{\lambda}_t$ mostra a relação de média e da variância, podendo apontar para a distribuição de Poisson se os pontos se espalharem pela função identidade ou para a distribuição Binomial Negativa se existir uma relação quadrática (ver em Ver Hoef & Boveng (2007)).

Christou *et al.* (2015) e Jung & Tremayne (2011) oferecem ferramentas para avaliar o desempenho preditivo de séries temporais de contagem, que foram originalmente propostas por Gneiting *et al.* (2007) e outros para dados contínuos e transferidas para independentes, mas não dados de contagem identicamente distribuídos por Czado *et al.* (2009). Essa ferramenta segue o *prequential principle* formulado por Dawid (1984), dependendo apenas das observações realizadas e as suas respectivas distribuições de predição. Temos que $P_t(y) = P(Y_t \leq y|\mathcal{F}_{t-1})$ é a função de distribuição acumulativa (f.d.a.), e $p_t(y) = P(Y_t = y|\mathcal{F}_{t-1})$ é a função de densidade de probabilidade, $y \in \mathbb{N}_0$, e temos $v_t = \sqrt{\mathrm{VAR}(Y_t|\mathcal{F}_{t-1})}$ é o desvio padrão da distribuição preditiva, que é uma distribuição de Poisson com média $\widehat{\lambda}_t$ ou a distribuição Binomial Negativa com média $\widehat{\lambda}_t$ e o coeficiente de dispersão $\widehat{\sigma}^2$.

A ferramenta para avaliar a calibração probabilística da distribuição preditiva (veja em Gneiting *et al.* (2007)) é a transformada da integral de probabilidade (*probability integral transform (PIT)*), que seguirá uma distribuição uniforme se a distribuição preditiva estiver correta. Para os dados de contagem Czado *et al.* (2009) definiu um valor de PIT não aleatório para o valor observado y_t e a distribuição preditiva $P_t(y)$ dado por

$$F_t(u|y) = \begin{cases} 0, & u \leq P_t(y-1) \\ \frac{u - P_t(y-1)}{P_t(y) - P_t(y-1)}, & P_t(y-1) < u < P_t(y) \\ 1, & u \geq P_t(y) \end{cases}$$

A média PIT é dado por

$$\overline{F}(u) = \frac{1}{n} \sum_{t=1}^{n} F_t(u|y_t), \quad 0 \leq u \leq 1\cdot$$

Para verificar se $\overline{F}(u)$ é uma f.d.a. de uma distribuição uniforme, Czado *et al.* (2009) propôs plotar um histograma com as caixas H, onde a caixa h tem a altura $f_j = \overline{F}(h/H) - \overline{F}((h-1)/H)$, $h = 1, \ldots, H$ (function `pit`). Por padrão, H é escolhido para ser 10. Uma forma em U indica subdispersão da distribuição preditiva, enquanto uma forma em U invertida indica superdispersão. Gneiting *et al.* (2007) afirmou que a cobertura empírica dos intervalos de previsão centrais, por exemplo, 90%, pode ser lida no histograma do PIT como a área sob as posições centrais de 90%.

A calibração marginal é definida como a diferença da média preditiva f.d.a. e o empírico f.d.a. das observações, ou seja,

$$\frac{1}{n} \sum_{t=1}^{n} P_t(y) - \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{\mathbf{y_t} \leq \mathbf{y}\} \tag{15}$$

para todos $y \in \mathbb{R}$. Na prática, plotamos a calibração marginal para valores y no intervalo das observações originais (Christou *et al.* 2015) (funcion `marcal`). Se as previsões de um modelo são apropriadas, a distribuição marginal das previsões assemelha-se à distribuição marginal das observações e a diferença 15 deve ser próxima de zero. Principais desvios do ponto zero para modelar deficiências.

Gneiting *et al.* (2007) mostrou que para a previsão ser ideal, a calibração avaliada por um histograma de PIT ou um gráfico de calibração marginal é uma condição necessária, mas não suficiente. Um bom modelo é aquele com a máxima nitidez entre todos os modelos suficientemente calibrados. A nitidez é a concentração da distribuição preditiva e pode ser medida pela largura dos intervalos de previsão. Uma avaliação simultânea de calibração e nitidez resumida em uma única pontuação numérica pode ser realizada por regras de pontuação adequadas (Gneiting *et al.* 2007). Denote uma pontuação para a distribuição

preditiva P_t e a observação y_t por $s(P_t, y_t)$. Um número de possíveis regras de pontuação adequadas é dado na Tabela 11. A pontuação média para cada modelo correspondente é dada por $\sum_{t=1}^{n} s(P_t, y_t)/n$. Cada uma das diferentes regras de pontuação adequada capta características diferentes da distribuição preditiva e sua distância aos dados observados (function `scoring`). Exceto para o escore de erro normalizado, o modelo com menor pontuação é preferível. O escore de erro quadrático médio é o único que não depende da distribuição e também é conhecido como erro de predição do quadrado médio. O escore de erro quadrático normalizado médio mede a variância dos resíduos de Pearson e está próximo de um se o modelo for adequado. O escore de Dawid-Sebastini é uma variante disso com um termo extra para penalizar a superestimação do desvio padrão.

Regra da Pontuação	**Abreviação**	**Definição**
escore do erro quadrático	`sqerror`	$(y_t - \lambda_t)^2$
escore do erro quadrático normalizado	`normsq`	$(y_t - \lambda_t)^2/v_t^2$
escore Dawid-Sebastiani	`dawseb`	$(y_t - \lambda_t)^2/v_t^2 + 2log(v_t$
escore logarítmico	`logarithmic`	$-log(p_t(y_t))$
escore quadrático (ou Brier)	`quadratic`	$-2_t(y_t) + \|p_t\|^2$
escore esférico	`spherical`	$-p_t(y_t)/\|p_t\|$
escore da probabilidade classificada	`rankprob`	$\sum_{y=0}^{\infty}(P_t(y) - \mathbb{I}(y_t \leq y))^2$

Tabela 11: Definições das regras de pontuação $s(P_t, y_t)$ e suas abreveaturas no pacote; $\|p_t\|^2 = \sum_{y=0}^{\infty} p_t^2(y)$.

3.5 Análise de intervenção

Liboschik *et al.* (2017) afirma que em muitas aplicações, ocorrem eventos extraordinários ou mudanças repentinas. Como por exemplo, o surto de uma epidemia em uma série temporal que conta o número semanal de pacientes infectados com uma doença em particular. Box & Tiao (1975) se referem a eventos especiais como intervenções. É interessante examinar o efeito de intervenções conhecidas, por exemplo, para julgar se uma mudança de política teve o impacto pretendido, ou para buscar efeitos de intervenção desconhecidos e encontrar explicações para eles a posteriori.

Fokianos & Fried (2010, 2012) modelam as intervenções que afetam a localização incluindo uma covariável determinística da forma $\delta^{t-\tau}\,\mathbb{I}(t \geq \tau)$, onde τ é o tempo de ocorrência e a taxa de decaimento δ é uma constante conhecida (funciton `interv_covariate`). Isto cobre vários tipos de intervenções para diferentes escolhas da constante δ: um efeito singular, ou seja, naquele ponto para $\delta = 0$ (spiky outlier), uma mudança com efeito de

decaimento exponencial na localização, isso quer dizer, ele altera, mas depois volta para $\delta \in (0, 1)$ (deslocamento transiente) e uma mudança permanente de localização, aumenta e permanece em outro nível para $\delta = 1$ (mudança de nível). Semelhante ao caso das covariáveis, o efeito de uma intervenção é essencialmente aditivo para o modelo linear e multiplicativo para o modelo log-linear. No entanto, a intervenção entra na dinâmica do processo e, portanto, seu efeito sobre o preditor linear não é puramente aditivo. Esse pacote **tscount** inclui métodos para testar tais efeitos de intervenção, desenvolvido por Fokianos & Fried (2010, 2012), adequadamente adaptados à classe de modelo mais geral descrita na Seção 3.1. Na Equação 16 temos o preditor linear de um modelo com s tipos de intervenções de acordo com os parâmetros $\delta_1, \ldots, \delta_s$ que ocorrem nos pontos de tempo $\tau_1, \ldots, \tau_s$.

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-j_l}) + \boldsymbol{\eta}^\top \mathbf{X}_t + \sum_{m=1}^{s} \omega_m \delta_m^{t-\tau_m} \mathbb{I}(t \geq \tau_m), \qquad (16)$$

em que ω_m, $m = 1, \ldots, s$ é o tamanho da intervenção. No momento de sua ocorrência, uma intervenção altera o nível da série temporal adicionando a magnitude ω_m, para um modelo linear como o Modelo 2, ou multiplicando o fator $exp\omega_m$, para um modelo log-linear como o Modelo 3. Nos próximos parágrafos, vai ser, brevemente resumido, os procedimentos de detecção de intervenção propostos.

O pacote **tscount** permite testar se as intervenções de certos tipos ocorrendo em determinados momentos, de acordo com o Modelo 16, têm efeito sobre as séries temporais observadas, ou seja, desejamos testar a hipótese $H_0 : \omega_1 = \ldots = \omega_s = 0$ contra a hipótese alternativa $H_1 : \omega_l \neq 0$ para algum $l \in 1, \ldots, s$. Isso é realizado aplicando um teste de score aproximado (function `interv_test`). Sob a estatística do teste $T_n(\tau_1, \ldots, \tau s)$ segue assintoticamente uma distribuição χ^2 com s graus de liberdade, assumindo algumas condições de regularidade (Fokianos & Fried 2010, Lemma 1).

Para testar se uma única intervenção de um certo tipo ocorrendo em um ponto de tempo desconhecido τ tem um efeito, o pacote emprega o máximo das estatísticas de teste de score $T_n(\tau)$ e determina um valor p por um procedimento de bootstrap paramétrico (funtion `interv_detect`). Se considerarmos um conjunto D de pontos no tempo em que a intervenção pode ocorrer, por exemplo, $D = 2, \ldots, n$, esta estatística de teste é

dada por $\tilde{T}_n = max_{\tau \in D} T_n(\tau)$. O procedimento de bootstrap pode ser calculado em vários núcleos simultaneamente (argument `parallel = TRUE`). O ponto temporal da intervenção é estimado como sendo o valor τ que maximiza esta estatística de teste. Geralmente, esse estimador apresenta uma grande variabilidade. Pode-se acelerar o cálculo das estatísticas do teste de bootstrap usando os parâmetros do modelo usados para geração dos exemplos de bootstrap em vez de estimar os mesmos para cada amostra de bootstrap (argument `final.control_bootstrap = NULL`). Isso resulta em um procedimento conservador, como observado por Fokianos & Fried (2012).

Se mais de uma intervenção for suspeita nos dados, mas nem seus tipos nem os pontos de tempo de suas ocorrências forem conhecidos, podemos utilizar um procedimento de detecção iterativo (function `interv_multiple`). Utiliza-se o conjunto de possíveis tempos de intervenção D como antes, e um conjunto de possíveis tipos de intervenção Δ, por exemplo, $\Delta = \{0, 0{,}8, 1\}$. Em um primeiro passo, a série temporal é testada para uma intervenção de cada tipo $\delta \in \Delta$ conforme descrito no parágrafo anterior, e os p valores são corrigidos para levar em conta os testes múltiplos pelo método de Bonferroni. Se nenhum dos p valores estiverem abaixo do nível de significância previamente especificado, o procedimento será interrompido e não identificará um efeito de intervenção. Caso contrário, o procedimento detecta uma intervenção do tipo correspondente ao menor p valor. No caso de p valores iguais, a preferência é dada às intervenções com $\delta = 1$, que são mudanças de nível, e depois àquelas com a mior estatística do teste. Em um segundo passo, o efeito de intervenção detectado é eliminado da série temporal e começa um novo procedimento até que não sejam detectados mais efeitos de intervenção. Finalmente, o Modelo 16 com todos os efeitos de intervenção detectados pode ser ajustado aos dados para estimar os tamanhos da intervenção e os outros parâmetros conjuntamente (que são geralmente diferentes do que quando estimados em etapas separadas). Note que a inferência estatística para este ajuste final do modelo deve ser feita com cuidado.

Nas aplicações práticas, a taxa de decaimento δ de um determinado efeito de intervenção é frequentemente desconhecida e precisa ser estimada. Como o parâmetro δ não é identificável quando o tamanho da intervenção ω é zero, sua estimativa é fora do padrão. Como sugerido por um revisor, a estimativa poderia ser realizada pelo resultado da probabilidade sobre este parâmetro. Para um único efeito de intervenção, isso poderia ser feito calculando o (quase) estimador MV de todos os outros parâmetros para uma dada taxa

de decaimento δ. Isso é repetido para todos $\delta \in \Delta$, em que Δ é um conjunto de possíveis taxas de decaimento e o valor que resulta no valor máximo da log-verossimilhança é escolhido (aplique a function `tsglm` repetidamente). Segundo Liboschik *et al.* (2017) esta abordagem afeta a validade da inferência estatística usual para os outros parâmetros.

Liboschik *et al.* (2016) estudam um modelo para efeitos de intervenção externa (modelado por efeitos covariáveis externos, Modelo 6 e a discussão relacionada) e comparam-no aos efeitos de intervenção interna estudados nas duas publicações referenciadas (argument `externo`).

4 Análise dos dados do número de atestados diários

Neste capítulo, a abordagem de MLG para séries temporais de dados de contagem discutida no Capítulo 3 será utilizada para analisar a série do número de atestados diários descrita no Capítulo 2. A Figura 7 apresenta o gráfico para a série considerada. Todas as análises foram feitas com o pacote **tscount** do **R**.

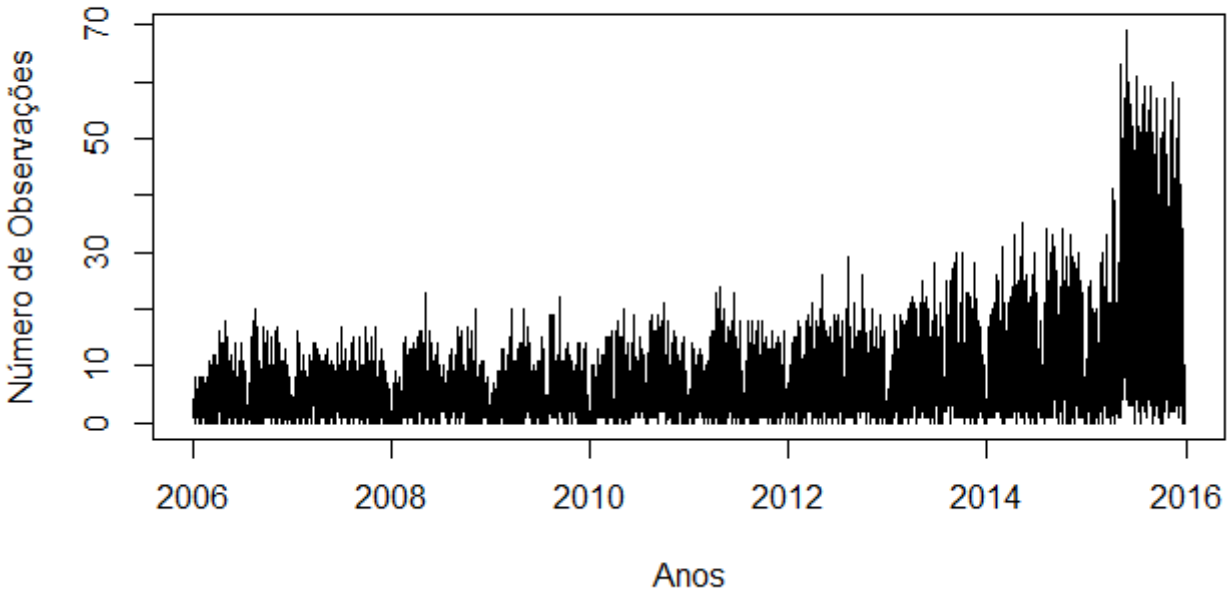


Figura 7: Número de observações (reportados diariamente) dos servidores públicos efetivos de um órgão público federal, de 2006 a 2015.

As distribuições condicionais consideradas são Poisson e Binomial Negativa, com função de ligação (`link = "log"`), pois ela permite o efeito negativo das covariáveis, assegurando que λ_t será sempre positiva. A covariável é determinística descrevendo tendência linear anual (function `linearTrend`), e é fornecida pelo argument `xreg`. Para definir os parâmetros p e q, que representam o número de defasagens que será considerado no modelo (argument `model = list(past_obs, past_mean`), foi feito um `grid` com todas as possíveis combinações de defasagens (ver código abaixo). Utiliza-se então o critério de informação Bayesiano (BIC) para selecionar as defasagens que fornecem os menores valores para o BIC melhor. Em particular, para os modelos baseados na Poisson e Binomial Negativa, os valores obtidos foram $p = 7$ e $q = 9$ (`past_obs = c(1:7), past_mean =`

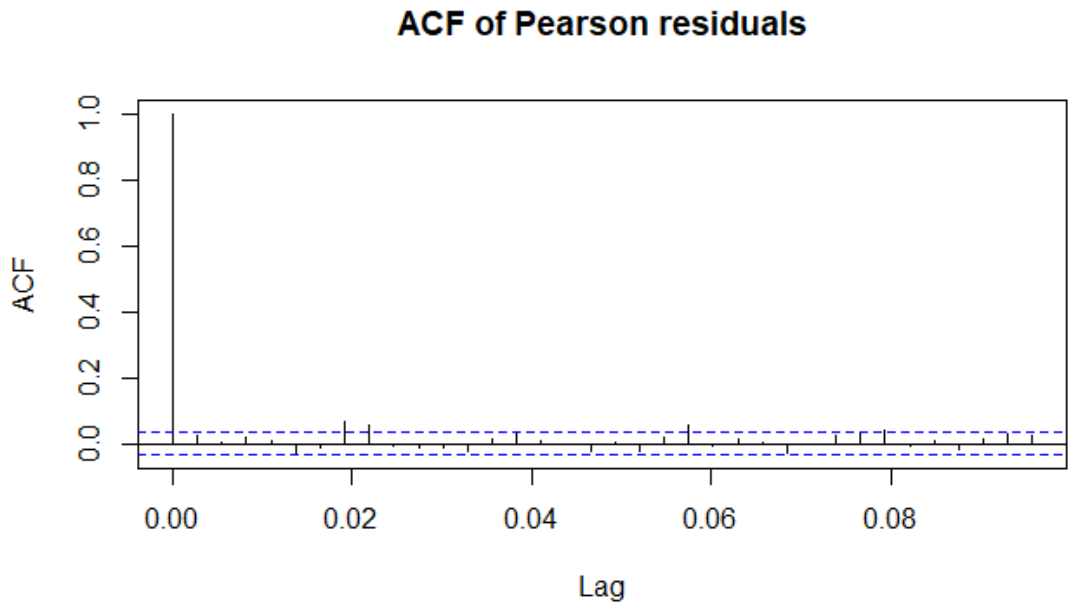
c(1:9)), de maneira que os modelos permitem regredir sobre 7 observações defasadas e 9 médias condicionais defasadas.

```
> regressors <- cbind(linearTrend = seq(along = Xts)/365.25)
> nd_pars <- expand.grid(pastobs = 1:12, pastmean = 1:12)
> nd_aic <- rep(0, nrow(nd_pars))
> for(i in seq(along = nd_aic))
+   nd_aic[i] <- AIC(tsglm(Xts,
+                          model = list(past_obs = c(1:nd_pars[i,1]),
+                                       past_mean = c(1:nd_pars[i,2])),
+                          link = "log", distr = "nbinom", xreg = regressors),
+                          k = log(length(Xts)))
> nd_pars[which.min(nd_aic),]
    pastobs pastmean
103       7        9
```

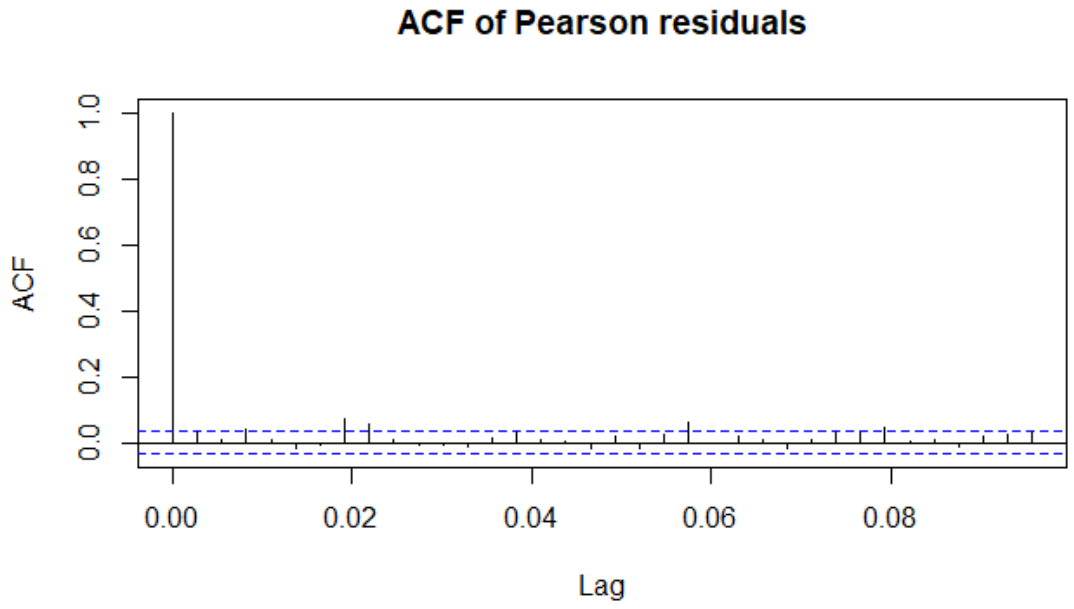
Temos duas possíveis distribuições para esse modelo, a distribuição de Poisson e a Binomial Negativa.

```
> fit_nbin <- tsglm(Xts, model = list(past_obs = c(1:7), past_mean = c(1:9)),
+              link = "log", distr = "nbinom", xreg = regressors )
> fit_pois <- tsglm(Xts, model = list(past_obs = c(1:7), past_mean = c(1:9)),
+              link = "log", distr = "poisson", xreg = regressors )
```

Para escolher qual distribuição se ajusta melhor ao modelo, existem algumas ferramentas como as citadas na análise de resíduos. A função de autocorrelação, ou comumente chamada de ACF do inglês, nos ajuda a caracterizar o desenvolvimento de uma série ao longo do tempo. Ela nos mostra o quão forte o valor observado hoje está correlacionado com os valores observados no passado, ou seja, esperamos que o ACF esteja dentro do intervalo de confiança. Como mostrado na Figura 8, os resíduos da resposta são muito parecidos para as duas distribuições condicionais, as ACF não exibem correlação serial ou sazonalidade que não tenha sido levada em consideração pelos modelos.



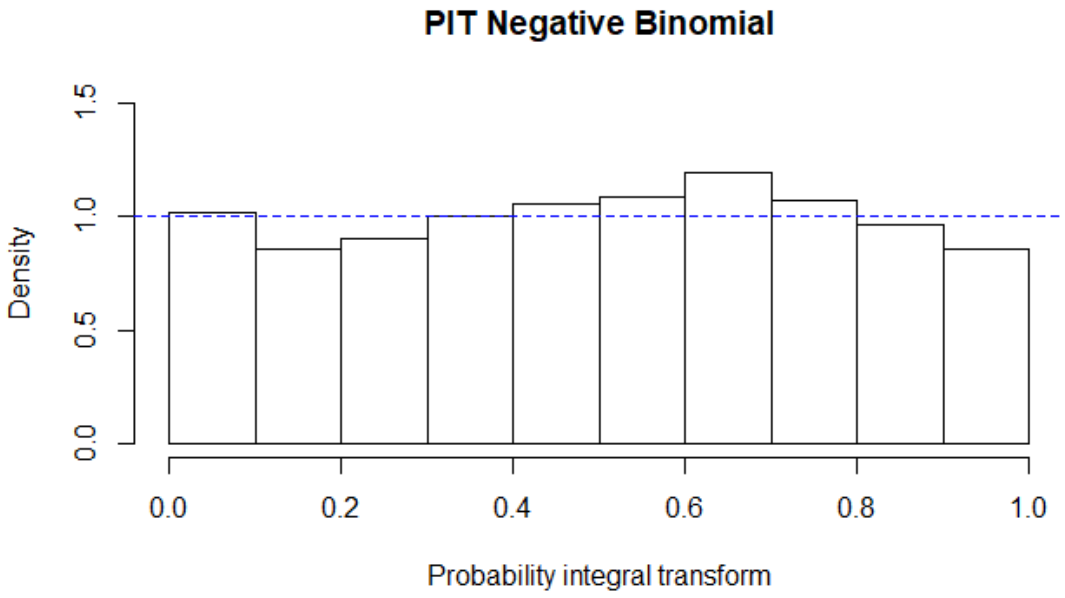
(a) Negative Binomial



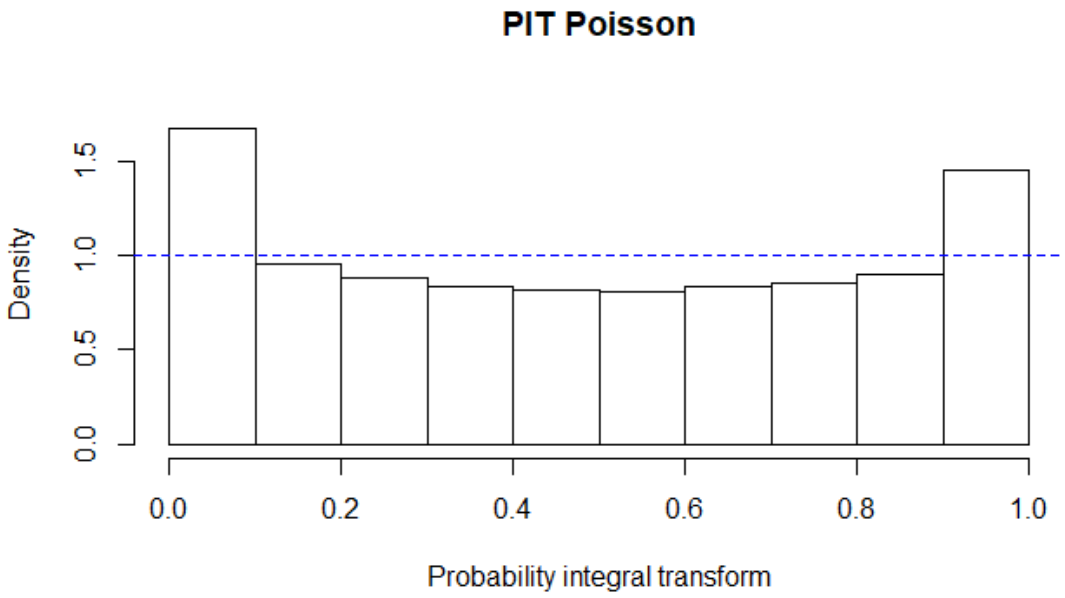
(b) Poisson

Figura 8: Função de Autocorrelação (ACF)

Na Figura 9 temos a PIT das duas distribuições, podemos observar que o histograma PIT da Poisson está em forma de U, indicando que essa não é a melhor distribuição para o modelo. Podemos observar que o histograma PIT da Binomial Negativa se aproxima melhor da distribuição uniforme, indicando que a calibração probabilística do modelo Binomial Negativo é satisfatória.



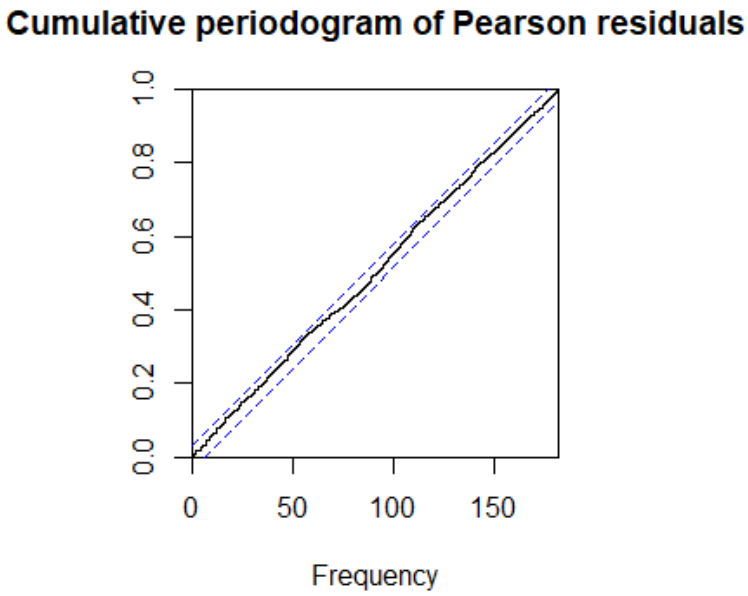
(a) Negative Binomial



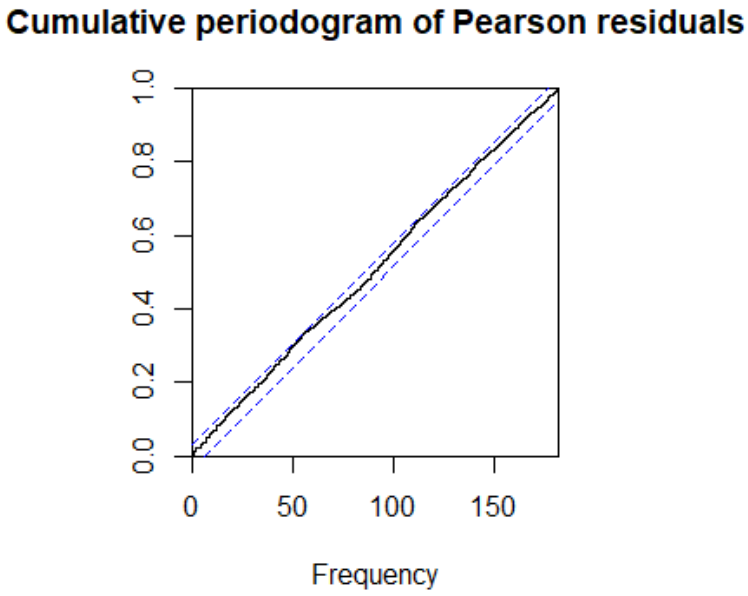
(b) Poisson

Figura 9: Transformada da Integral da Probabilidade (PIT)

O teste do periodograma acumulado é utilizado como uma ferramente para verificar a aleatoriedade dos resíduos após o ajuste do modelo. Podemos observar na Figura 10 que a série dos resíduos padronizados se aproxima a do ruído branco, ou seja, o modelo proposto é válido.



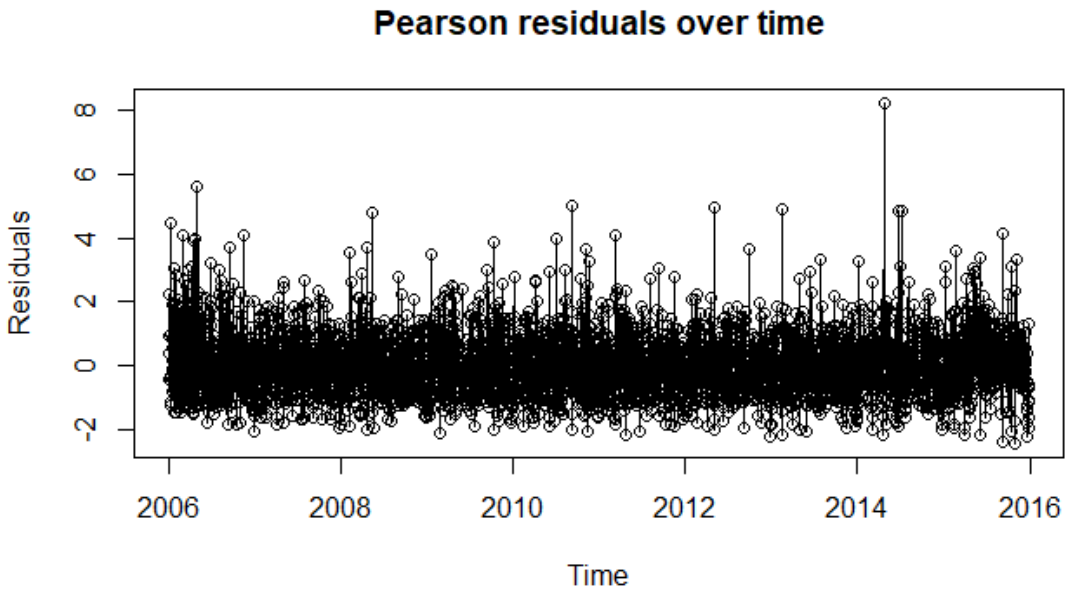
(a) Negative Binomial



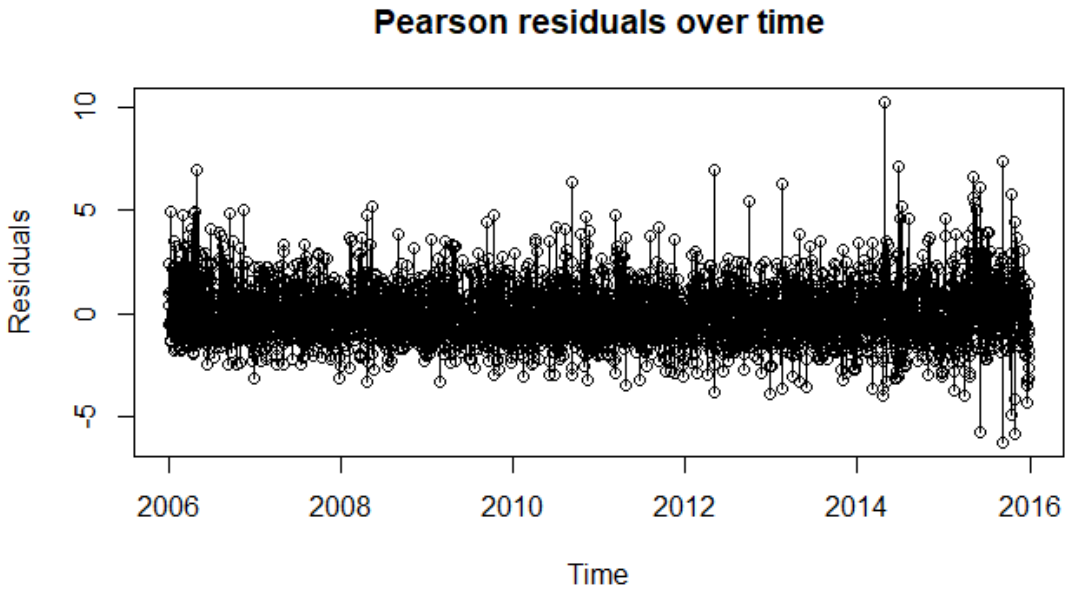
(b) Poisson

Figura 10: Periodograma Acumulado dos Resíduos de Pearson

A Figura 11 nos mostra o gráfico dos resíduos de Pearson ao longo do tempo, um modelo que se ajusta bem aos dados deve ter uma distribuição de pontos aleatoriamente dispersos. Como pode-se observar, ambas as distribuições apresentam comportamento aleatório, ou seja, não apresentam nenhum padrão fixo ou tendência.



(a) Negative Binomial



(b) Poisson

Figura 11: Gráfico dos Resíduos de Pearson ao longo do tempo

Como último critério, temos as regras de pontuação para as duas distribuições, e todas as regras consideradas são favoráveis à distribuição Binomial Negativa.

```
> rbind(Poisson = scoring(fit_pois), BegBin = scoring(fit_nbin))
        logarithmic  quadratic  spherical rankprob   dawseb    normsq  sqerror
Poisson    2.669019 -0.1186341 -0.3277678 2.084174 3.599334 1.9183772 19.19753
```

```
BegBin      2.514922 -0.1219064 -0.3296856 2.065974 3.309107 0.9950712 19.19753
```

A variável resposta, o número de atestados diários, tem uma variância grande, ou seja, tem-se uma superdispersão dos dados, característica que a distribuição Poisson não consegue acomodar, uma vez que nesse modelo $\text{VAR}(Y_t|\mathcal{F}_{t-1}) = \mathbb{E}(Y_t|\mathcal{F}_{t-1})$. Desse modo, baseado nesse fato, nos gráficos anteriores e nos resultados obtidos pelas regras de pontuação, o melhor modelo é aquele baseado na Binomial Negativa, com as respectivas estimativas dadas por:

```
    > summary(fit_nbin)
Call:
tsglm(ts = Xts, model = list(past_obs = c(1:7), past_mean = c(1:9)),
    xreg = regressors, link = "log", distr = "nbinom")


Coefficients:
             Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)   0.00299    0.04487    -0.0850     0.0909
beta_1        0.22157    0.01847     0.1854     0.2578
beta_2        0.21000    0.01857     0.1736     0.2464
beta_3        0.20042    0.01715     0.1668     0.2340
beta_4        0.19095    0.01704     0.1576     0.2243
beta_5        0.19337    0.01684     0.1604     0.2264
beta_6        0.19107    0.01663     0.1585     0.2237
beta_7        0.25237    0.01645     0.2201     0.2846
alpha_1      -0.42819    0.08789    -0.6005    -0.2559
alpha_2      -0.45613    0.07983    -0.6126    -0.2997
alpha_3      -0.23172    0.01957    -0.2701    -0.1934
alpha_4      -0.22136    0.01958    -0.2597    -0.1830
alpha_5      -0.22826    0.01936    -0.2662    -0.1903
alpha_6      -0.22383    0.01958    -0.2622    -0.1854
alpha_7       0.72640    0.01942     0.6883     0.7645
alpha_8       0.18092    0.07673     0.0305     0.3313
alpha_9       0.21853    0.07094     0.0795     0.3576
linearTrend   0.03771    0.00921     0.0197     0.0558
```

```
sigmasq       0.13003         NA         NA         NA
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: log
Distribution family: nbinom (with overdispersion coefficient 'sigmasq')
Number of coefficients: 19
Log-likelihood: -9184.495
AIC: 18406.99
BIC: 18524.85
QIC: 19595.54
```

Na saída acima, tem-se o coeficiente de superdispersão σ^2, que está relacionado ao parâmetro de dispersão ϕ da distribuição Binomial Negativa, $\phi = 1/\sigma^2$. Logo, o modelo ajustado para o número de atestados médicos Y_t no tempo t é dado por $Y_t|\mathcal{F}_{t-1} \sim \text{BinNeg}(\lambda_t, 7{,}69)$ com

$$\begin{aligned} log(\lambda_t) &= 0{,}003 + 0{,}221Y_{t-1} + 0{,}21Y_{t-2} + 0{,}2Y_{t-3} + 0{,}19Y_{t-4} + 0{,}193Y_{t-5} \\ &\quad +0{,}191Y_{t-6} + 0{,}252Y_{t-7} - 0{,}428\lambda_{t-1} - 0{,}4566\lambda_{t-2} - 0{,}232\lambda_{t-3} \\ &\quad -0{,}221\lambda_{t-4} - 0{,}228\lambda_{t-5} - 0{,}224\lambda_{t-6} + 0{,}726\lambda_{t-7} + 0{,}181\lambda_{t-8} \\ &\quad +0{,}218\lambda_{t-9} + 0{,}038t/365, \qquad t = 1, \ldots, 3652\cdot \end{aligned}$$

A tendência linear no modelo acima pode ser interpretada como um aumento anual de 0.038 no número de atestados, ou seja, em média espera-se um aumento de 3.8% no número de atestados a cada ano.

Com base na série do número de atestados diários até novembro de 2015, pode-se prever o mesmo para dezembro de 2015. A tabela 12 apresenta o valor real do número de atestados diários para dezembro de 2015, os valores preditos e os intervalos de predição. De acordo com o modelo de distribuição binomial negativa, os intervalos de predição são calculados para garantir uma taxa de cobertura global de 90%.

```
Xts_2015 <- window(Xts, start = 2006, end = c(2015,340))
regressors_2015 <- cbind(linearTrend = seq(along = Xts_2015)/365)

fit_pred <- tsglm(Xts_2015, model = list(past_obs = c(1:7), past_mean = c(1:9)),
```

```
        link = "log", distr = "nbinom",xreg = regressors_2015 )

Xts_2015_rest <- window(Xts, start = c(2015,341),end=c(2016,2))
regressors_2015_rest <- cbind(linearTrend = seq(along = Xts_2015_rest)/365)

predic_rest <- predict(fit_pred, n.ahead = 27, level = 0.9, global = TRUE, B = 2000
       newxreg = regressors_2015_rest)\$pred
```

Tabela 12: Valores reais, valores preditos e intervalos de predição para a série do número de atestados diários.

Valor Real	Valor Predito	Intervalo Inferior	Intervalo Superior
2	2	0	7
1	2	0	8
57	37	6	87
26	28	4	72
26	29	6	75
34	35	4	88
33	29	4	74
3	2	0	6
4	2	0	7
42	32	5	81
34	24	3	62
22	24	3	59
25	30	5	79
34	25	3	60
0	2	0	5
1	2	0	6
14	28	3	65
15	21	3	55
5	21	1	53
1	27	3	68
1	23	2	63
0	2	0	6
2	2	0	6
10	25	2	60
6	19	2	44
7	19	1	49
3	25	2	57

Por final, testa-se se houve uma mudança abrupta no número de atestados após a implementação do ponto eletrônico em maio de 2015. Ou seja, tem-se as seguintes hipóteses,

H_0 = Não houve intervenção e H_1 = Houve intervenção. O teste de intervenção descrito no Seção 3.5 é aplicado:

```
    > interv_test(fit_nbin, tau = 3408, delta = 1, est_interv = TRUE)

Score test on intervention(s) of given type at given time

Chisq-Statistic: 14.2518 on 1 degree(s) of freedom, p-value: 0.000159908

Fitted model with the specified intervention:

Call:
tsglm(ts = fit$ts, model = model_extended, xreg = xreg_extended,
    link = fit$link, distr = fit$distr)

Coefficients:
(Intercept)       beta_1       beta_2       beta_3       beta_4       beta_5
    0.73765      0.21552      0.20776      0.20236      0.18921      0.19244
     beta_6       beta_7      alpha_1      alpha_2      alpha_3      alpha_4
    0.19107      0.24356     -0.47967     -0.51436     -0.30923     -0.29755
    alpha_5      alpha_6      alpha_7      alpha_8      alpha_9  linearTrend
   -0.30335     -0.29933      0.65734      0.16128      0.20161      0.06222
    interv_1
    0.71699

Overdispersion coefficient 'sigmasq' was estimated to be 0.119815.
```

Com um p-valor de aproximadamente 0.00016, rejeita-se a hipótese nula de não intervenção a um nível de significância de 5%. Esse resultado corrobora então, a evidência de mudança no número de atestados após a implementação do ponto eletrônico, encontrada na análise descritiva realizada no Capítulo 2.

5 Conclusão

Nesse trabalho foi feita uma análise descritiva do número de dias de afastamento e do número de atestados, ambos diários. Vimos que os seguintes fatores podem afetar essas variáveis: tipo de doença (CID), gênero, idade, departamento, tipo de vínculo.

Observamos também, que a série do número de atestados por dia ao longo dos anos é crescente, e existe uma mudança repentina nessa série, que podemos chamar de intervenção - implantação do ponto eletrônico. A observação desta variação brusca gerou nosso interesse em perseguir uma análise mais aprofundada, que permitisse melhor entendimento da situação, possibilitando ações benéficas futuras neste ambiente laboral.

Utilizou-se modelos lineares generalizados para séries temporais de contagem com distribuições condicionais Poisson e Binomial Negativa para modelar a série do número de atestados diários. Os resultados indicaram que o modelo baseado na Binomial Negativa é mais adequado. Foi feita uma previsão para dezembro de 2015, e os valores preditivos se aproximaram dos valores reais. Ainda, atavés de uma análise de intervenção, observou-se que a hipótese nula de nenhuma mudança abrupta na série devido a inserção do ponto eletrônico pode ser rejeita a um nível de 5%.

Os resultados obtidos através desta análise, podem servir como ponto de partida para ações com o objetivo de melhorar a saúde do trabalhador e até diminuir os afastamentos por motivo de saúde, reduzindo o custo à sociedade.

Referências Bibliográficas

Ahmad, Ali, & Francq, Christian. 2016. Poisson QMLE of count time series models. *Journal of Time Series Analysis*, **37**(3), 291–314.

Box, George EP, & Tiao, George C. 1975. Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, **70**(349), 70–79.

Christou, Vasiliki, & Fokianos, Konstantinos. 2014. Quasi-Likelihood Inference for Negative Binomial Time Series Models. *Journal of Time Series Analysis*, **35**(1), 55–78.

Christou, Vasiliki, Fokianos, Konstantinos, *et al.* . 2015. Estimation and testing linearity for non-linear mixed poisson autoregressions. *Electronic Journal of Statistics*, **9**(1), 1357–1377.

Czado, Claudia, Gneiting, Tilmann, & Held, Leonhard. 2009. Predictive model assessment for count data. *Biometrics*, **65**(4), 1254–1261.

Dawid, A Philip. 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 278–292.

Douc, Randal, Doukhan, Paul, & Moulines, Eric. 2013. Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications*, **123**(7), 2620–2647.

Fahrmeir, Ludwig. 2001. Multivariate Statistical Modelling Based on Generalized Linear Models.

Ferland, René, Latour, Alain, & Oraichi, Driss. 2006. Integer-Valued GARCH Process. *Journal of Time Series Analysis*, **27**(6), 923–942.

Fokianos, Konstantinos, & Fried, Roland. 2010. Interventions in INGARCH processes. *Journal of Time Series Analysis*, **31**(3), 210–225.

Fokianos, Konstantinos, & Fried, Roland. 2012. Interventions in log-linear Poisson autoregression. *Statistical Modelling*, **12**(4), 299–322.

Fokianos, Konstantinos, & Tjøstheim, Dag. 2011. Log-linear Poisson autoregression. *Journal of Multivariate Analysis*, **102**(3), 563–578.

Fokianos, Konstantinos, Rahbek, Anders, & Tjøstheim, Dag. 2009. Poisson autoregression. *Journal of the American Statistical Association*, **104**(488), 1430–1439.

Gneiting, Tilmann, Balabdaoui, Fadoua, & Raftery, Adrian E. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 243–268.

Heinen, Andréas. 2003. Modelling time series count data: an autoregressive conditional Poisson model.

Hilbe, Joseph M. 2011. *Negative binomial regression*. Cambridge University Press.

Jung, Robert C, & Tremayne, AR. 2011. Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, **95**(1), 59–91.

Kedem, Benjamin, & Fokianos, Konstantinos. 2002. *Regression models for time series analysis*.

Liboschik, Tobias, Kerschke, Pascal, Fokianos, Konstantinos, & Fried, Roland. 2016. Modelling interventions in INGARCH processes. *International Journal of Computer Mathematics*, **93**(4), 640 – 657.

Liboschik, Tobias, Fokianos, Konstantinos, & Fried, Roland. 2017. tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, **85**(5), 1 – 51.

Nelder, JA, & Wedderburn, RWM. 1972. Generalizedlinear models-JR Statist. Soc. A, 135: 370-384. Nelder370135J. R. *Statist. Soc A*, **1972**.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Venables, WN, & Ripley, BD. 2002. Random and mixed effects. *Pages 271–300 of: Modern applied statistics with S*. Springer.

Ver Hoef, Jay M, & Boveng, Peter L. 2007. Quasi-poisson Vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data? *Ecology*, **88**(11), 2766–2772.

Woodard, Dawn B, Matteson, David S, Henderson, Shane G, *et al.* . 2011. Stationarity of generalized autoregressive moving average models. *Electronic Journal of Statistics*, **5**, 800–828.