EMERSON URIZZI CERVI

Manual de

MÉTODOS QUANTITATIVOS

PARA INICIANTES EM CIÊNCIA POLÍTICA

Vol. 2

CPOP
UFPR

Manual de Métodos Quantitativos para iniciantes em Ciência Política

Volume 2

EMERSON URIZZI CERVI

MANUAL DE MÉTODOS QUANTITATIVOS PARA INICIANTES EM CIÊNCIA POLÍTICA

VOLUME 2

CURITIBA
2019

Dados Internacionais de Catalogação da Publicação
Fundação Biblioteca Nacional

C419m Cervi, Emerson U.
Manual de métodos quantitativos para iniciantes em Ciência Política – Vol. 2 / Emerson Urizzi Cervi - Curitiba: CPOP, 2019. (1ª edição).
314 p.

ISBN 978-85-915195-5-2

1. Pesquisa – Métodos Quantitativos. 2. Ciência Política. I. Título.

CDD-320
CDU: 001.8:303

Editorado pelo Grupo de Pesquisa em Comunicação Política e Opinião Pública - CPOP

Edição do Autor
Emerson Urizzi Cervi

Capa, Planejamento Gráfico e Diagramação
Fernanda Cavassana de Carvalho

Grupo de Pesquisa em Comunicação Política e Opinião Pública – CPOP
Programa de Pós-graduação em Ciência Política – PPGCP-UFPR
Universidade Federal do Paraná – Campus Reitoria
Rua General Carneiro, nº 460 – Ed. Dom Pedro I, 5º andar.
CEP: 80.060-000. Curitiba – Paraná – Brasil

www.cpop.ufpr.br | www.facebook.com/cpopufpr | nucleocpopufpr@gmail.com

SUMÁRIO

APRESENTAÇÃO 9

CAPÍTULO I – COEFICIENTE DE ASSOCIAÇÃO E ANÁLISE DE RESÍDUOS EM TABELAS DE CONTINGÊNCIA 11

1.1 Coeficiente Cramer's V para força de associação do χ^2 12

1.2 Coeficiente Delta (Δ) para diferenças de F_O e F_E 15

1.3 Testes de associação entre categorias de variáveis nominais em tabelas de contingência (Resíduos Brutos e Resíduos Padronizados) 19

1.3.1 Cálculo dos Resíduos Brutos (R_B) 21

1.3.2 Cálculo dos Resíduos Padronizados (R_P) 24

1.3.3 Resíduos Padronizados para análises temporais 27

1.4 Referências bibliográficas do Capítulo I 32

1.5 Exercícios propostos do Capítulo I 33

Anexo do Capítulo I 34

CAPÍTULO II – TESTE DE ASSOCIAÇÃO PARA TABELAS QUÁDRUPLAS E PARA VARIÁVEIS ORDINAIS 35

2.1 Teste Q de Yule (Q_{XY}) 36

2.1.1 Teste de Independência Q de Yule (Q_{XY}) 39

2.1.2 Cálculos Adicionais: proporções de pares consistentes e pares inconsistentes 44

2.1.3 Cálculos Adicionais: validade para inferências 45

2.1.4 Intervalo de Confiança para o Teste de Correlação Q_{XY} 46

2.1.5 Coeficiente Q_{XY} para três variáveis ($Q_{XY:T}$) 49

2.2 Teste com variáveis ordinais - Coeficiente Gama (G) 59

2.3 Referências bibliográficas do Capítulo II 64

2.4 Exercícios propostos do Capítulo II 65

CAPÍTULO III – ANÁLISE DE DADOS CATEGÓRICOS 67
3.1 Análise de dados Categóricos 68
3.2 Teste de confiabilidade para indicadores estatísticos 70
3.3 Testes estatísticos para associações bi e multivariados 75
3.3.1 Análise de Correspondência Canônica (ACC) 77
3.3.2 Teste de Múltipla Correspondência 82
3.3.3 Análise de Componentes Principais (PCA) 86
3.3.4 Análises de agrupamentos (*Cluster*) 90
3.4 Referências bibliográficas do Capítulo III 95
3.5 Exercícios propostos do Capítulo III 96
Anexos do Capítulo III 98

CAPÍTULO IV – ANÁLISE DE CONTEÚDO APLICADA A REDES SOCIAIS ONLINE 101
4.1 Histórico da Análise de Conteúdo 102
4.2 Etapas da Análise de Conteúdo aplicada a textos políticos 106
4.3 Descrição da proposta de análise em duas etapas com método Reinert 108
4.4 O método Reinert na análise de conteúdo de redes sociais online 110
4.5 Uma comparação com o método tradicional de classificar textos políticos 115
4.6 Análises das classificações a partir da tematização automatizada 119
4.7 Referências bibliográficas do Capítulo IV 126
4.8 Exercícios propostos do Capítulo IV 128

CAPÍTULO V – ANÁLISE DE REDES SOCIAIS 129
5.1 Conceituando Análise de Redes Sociais (ARS) 130
5.2 Componentes da ARS 132
5.3 Etapas para análise de redes sociais 136
5.4 Redes de financiamento de empresas a partidos políticos no Brasil 139
5.5 Referências bibliográficas do Capítulo V 145
5.6 Exercícios propostos do Capítulo V 146

CAPÍTULO VI – TESTES DE CORRELAÇÃO 147
6.1 Correlação Linear Simples 147
6.2 Aplicação dos testes de correlação para amostras 157
6.3 Coeficiente Linear de Determinação e de Alienação 159
6.4 Pressupostos a serem respeitados em análises de correlação 161
6.4.1 Transformações de dados para normalização de distribuições 164
6.5 Aplicação da correlação de Pearson e outros coeficientes no *RCommander* 165
6.6 Referências bibliográficas do Capítulo VI 169
6.7 Exercícios propostos do Capítulo VI 170

CAPÍTULO VII – PRINCÍPIOS DOS TESTES DE REGRESSÃO 171
7.1 Começando pelo início: regressão linear simples 173
7.2 Fórmula da Regressão Linear 176
7.3 Erro da reta de regressão (rms) – análise de resíduos 184
7.4 A estatística "t" e os testes complementares de ajustamento do modelo 187
7.5 Regressão Binária Logística 192
7.6 Referências bibliográficas do Capítulo VII 198
7.7 Exercícios propostos do Capítulo VII 199

CAPÍTULO VIII – ANÁLISE DE TRAJETÓRIA (*path analysis*) 201
8.1 Princípios da análise de trajetória 202
8.2 Componentes do modelo de análise de trajetória 205
8.3 Aplicação do modelo de análise de trajetória 208
8.4 Referências bibliográficas do Capítulo VIII 217
8.5 Exercícios propostos do Capítulo VIII 219
Anexos do Capítulo VIII 220

CAPÍTULO IX – ANÁLISE GEOGRÁFICA 221
9.1 Princípios e objetivos da Análise Geográfica 222
9.2 Bases de dados e *softwares* para Análise Geográfica 226
9.3 Mapas coropléticos no Geoda 228
9.3.1 Mapa Quantil 229
9.3.2 Mapa Percentil 230
9.3.3 *Box Map* 230
9.3.4 Mapa de desvio padrão 231
9.3.5 Mapa de valores únicos 232
9.3.6 Mapa de quebras naturais 233
9.3.7 Mapa com intervalos iguais 234
9.3.8 Mapa de razão de chance (*Excess Risk*) 235
9.4 Estatísticas básicas em análises geográficas descritivas 237
9.4.1 Autocorrelação espacial global com coeficiente I de *Moran* 238
9.4.2 Coeficiente LISA para *clusters* geográficos 240
9.4.3 Testes de regressão linear para unidades espaciais no *Geoda* 242
9.5 Referências bibliográficas do Capítulo IX 248
9.6 Exercícios propostos do Capítulo IX 250

CAPÍTULO X – ANÁLISE DE ANÁLISE DE SÉRIES TEMPORAIS 251
10.1 Fundamentos 252
10.2 Médias móveis 255
10.3 Funções de Autocorrelação (FAC) e Autocorrelação Parcial (FACP) 260
10.4 Teste Autoregressivo com médias móveis integradas (ARIMA) 262
10.5 Teste para Raízes Unitária 267
10.6 Análise multivariada no tempo (efeitos de intervenção e de transferência) 269
10.7 Referências bibliográficas do Capítulo X 276
10.8 Exercícios propostos do Capítulo X 277
Anexos do Capítulo X 278

ADENDO I - GABARITO DE RESPOSTAS AOS EXERCÍCIOS PROPOSTOS

Respostas às questões do Capítulo I 281
Respostas às questões do Capítulo II 282
Respostas às questões do Capítulo III 285
Respostas às questões do Capítulo IV 290
Respostas às questões do Capítulo V 299
Respostas às questões do Capítulo VI 300
Respostas às questões do Capítulo VII 302
Respostas às questões do Capítulo VIII 305
Respostas às questões do Capítulo IX 306
Respostas às questões do Capítulo X 309

APRESENTAÇÃO

Como o próprio nome indica, este Manual de Métodos Quantitativos para Iniciantes é uma continuidade do volume 1. Assim, é altamente recomendável que o aluno conheça os conceitos discutidos no volume anterior, que pode ser acessado em http://www.cpop.ufpr.br/publicacoes/metodos-quantitativos-para-iniciantes-v1. Como todo manual, trata-se de um material para estudos. Aqui, em cada capítulo é apresentada uma técnica para análise quantitativa de dados em função dos objetivos do pesquisador. Portanto, não se espera que alguém faça a leitura linear, do início ao fim do livro. Deve-se identificar qual capítulo trata da técnica de interesse para o seu trabalho e ir diretamente a ele.

É importante repetir nesta apresentação o que já foi destacado no volume anterior: trata-se de um manual introdutório para iniciantes, pensado para o nível de graduação. Para quem já é iniciado nas técnicas ou procura aprofundar conceitos e teorias sobre testes e estatísticas, esta obra não é recomendada.

O livro começa com as técnicas de análise descritiva básica a partir de tabelas de contingência, análise de dados categóricos e testes de diferenças de médias. Depois passa para os testes de correlação e apresenta uma introdução aos conceitos de regressão linear. Os últimos três capítulos do livro apresentam adaptações da técnica básica de regressão linear para finalidades específicas: análise de trajetória, análises espaciais e análise de séries temporais.

Seguindo o modelo do volume 1, aqui também são utilizados exclusivamente

softwares de código aberto, não sendo necessária a utilização de nenhum programa proprietário. A maior parte dos programas usados aqui é vinculada ao pacote estatístico R ou são *plug-ins* dele. Por se tratar de um manual introdutório a alunos de graduação, evita-se ao máximo o uso do pacote R diretamente. Sempre que possível, a opção é por uma interface mais "amigável" que a original para facilitar a vida dos não iniciados em programação computacional. Assim como no primeiro volume, ao final de cada capítulo são apresentadas as referências bibliográficas usadas para discussão de cada técnica e são propostos alguns exercícios para aprofundamento de aprendizagem. Todos os bancos de dados usados no livro e nos exercícios estão disponíveis "na nuvem" para *download*. O objetivo é permitir ao leitor que avance por conta própria em suas análises.

Este livro não existiria sem a contribuição dos alunos que integram o grupo de pesquisas em Comunicação Política e Opinião Pública (www.cpop.ufpr.br) da Universidade Federal do Paraná (UFPR). São eles que me estimulam a pensar formas didáticas de apresentar ferramentas e discutir técnicas de análise empírica para alunos de graduação. Eles também participaram diretamente da coleta e formação de alguns bancos de dados utilizados aqui. O manual é resultado de cursos de metodologia ministrados por mim ao longo dos últimos anos. Dentre os alunos do grupo de pesquisa, um agradecimento especial a Fernanda Cavassana, que além de ser a responsável direta pelo projeto gráfico e editoração dos dois volumes, também estabeleceu prazos e me cobrou o cumprimento dos mesmos. Enfim, feita a apresentação do volume II do manual, sinta-se à vontade para explorá-lo e para informar possíveis inconsistências ou erros ao longo do texto.

Bons estudos!

Curitiba, janeiro de 2019.

Capítulo I

Coeficiente de associação e análise de resíduos em tabelas de contingência

Identificar associações entre duas ocorrências é um desafio muito grande, mas que, infelizmente, costuma ter sua importância minimizada.

O atual capítulo é uma continuidade do capítulo VII do volume I do manual. Então, em vários momentos são feitas referências diretas a conteúdos que já foram apresentados naquele livro, que tem por objetivo servir como fundamento para os testes que serão apresentados a partir daqui. Para acessá-lo, basta clicar no link que se encontra no rodapé desta página[1]. Aqui, começo apresentando o teste adequado para medir o grau de associação entre duas variáveis após a realização do teste de independência de médias qui-quadrado (χ^2). É o coeficiente Cramer's V. Em seguida, são apresentadas formas estatísticas de medir a relação entre pares de categorias das variáveis. São o coeficiente Delta (Δ) e os Resíduos Padronizados. Ao final do capítulo, são apresentados exercícios e as referências bibliográficas citadas.

[1] Volume I do Manual de Métodos Quantitativos para Iniciantes em Ciência Política disponível em: http://www.cpop.ufpr.br/publicacoes/metodos-quantitativos-para-iniciantes-v1

1.1 Coeficiente Cramer's V para força de associação do χ^2

Uma vez identificado o valor do coeficiente de χ^2, como apresentado no capítulo VII do volume I, um erro muito comum é tirar conclusões sobre a magnitude da relação entre as duas variáveis apenas a partir desse coeficiente ou do seu nível de significância. Quando o coeficiente é alto, podemos dizer que as variações analisadas não são independentes, elas variam com algum grau de dependência uma da outra. Se considerarmos ainda os graus de liberdade do teste, podemos identificar o limite crítico do Intervalo de Confiança e se o coeficiente ficar abaixo desse limite pode-se dizer que a dependência das variações é forte o suficiente para ser extrapolada a toda a população - caso estejamos trabalhando com uma amostra. Isso porque a significância do teste depende no número de casos (graus de liberdade). Quanto maior a amostra ou população testada, maiores as chances do resultado ser estatisticamente significativo (Pereira, 2004). No entanto, nenhum dos coeficientes tratados até aqui é indicativo da relação entre duas variáveis categóricas para a magnitude ou força da associação entre elas.

Para tanto, existem coeficientes específicos que medem a força do efeito da variação de uma variável sobre a variação de outra. Quando estamos usando o teste de χ^2, para independência de médias, os testes de associação indicados são o coeficiente *Phi* e o coeficiente Cramer's V. Para identificar a magnitude do efeito (***effect size***) em testes de χ^2 em que se rejeita a hipótese nula, usa-se o coeficiente *Phi* para os casos de tabelas quádruplas (2x2), mesma indicação do Q de Yule, ou o coeficiente Cramer's V para tabelas maiores (Ln x Cn). Por agora, não trataremos do coeficiente *Phi*, pois dedicamos mais adiante um capítulo específico para o Q-yule, teste indicado para o mesmo tipo de variável a que se aplica o *Phi*.

A seguir, apresentamos como calcular o coeficiente de magnitude do efeito Cramer´s V para testes de χ^2. A indicação é calcular o Cramer's V apenas quando o coeficiente χ^2 for estatisticamente significativo, caso contrário, a magnitude do efeito será muito baixa ou nula.

A leitura dos resultados do coeficiente V é equivalente à de um coeficiente de correlação de Pearson. Ele indica qual a força da associação direta entre o conjunto

das categorias das duas variáveis testadas. A fórmula é a seguinte:

$$v = \sqrt{\frac{\chi^2}{N.(k-1)}}$$

Onde:

χ^2 = coeficiente qui-quadrado.

N = número de casos

K = número de categorias de uma das variáveis testadas. Utiliza-se sempre o menor número de categorias, independente de estar nas linhas ou nas colunas.

A leitura do resultado do Cramer's V é similar à de um coeficiente de correlação de Pearson. Sendo assim, o quadrado do seu valor (V^2) nos indica qual a proporção da variância da relação que é explicada pelo χ^2. Por exemplo, um V = 0,12, se elevado ao quadrado e multiplicado por 100 nos indicará qual o percentual de variância explicada. No caso, apenas 1,44% de variância explicada pelo χ^2, o que é um percentual bastante baixo em termos gerais. Quando a tabela testada é quádrupla (2x2) a fórmula do teste Cramer's V iguala-se à do *Phi*, pois nesse caso o número de categorias menos um sempre será um. Assim, a fórmula é reduzida a raiz quadrada de χ^2 dividido pelo número de casos.

Para exemplificar o uso do coeficiente Cramer's V, vamos fazer o teste de χ^2 para a associação entre sexo de vereadores eleitos em 2016 nas eleições municipais brasileiras (homem ou mulher) e região do País (norte, nordeste, centro-oeste, sudeste e sul). A hipótese nula defende que não há diferença nas variações entre as duas variáveis e que, portanto, homens e mulheres distribuem-se igualmente entre os eleitos nas cinco regiões do País. Nunca é demais lembrar que estamos testando a associação entre duas variáveis categóricas nominais. Aqui, estão considerados apenas os vereadores eleitos que indicaram sexo Homem ou Mulher no registro de candidatura no TSE em 2016.

Tabela 1.1.1. Distribuição dos vereadores eleitos por sexo e região do País em 2016

REGIÃO	HOMEM	MULHER	TOTAL
NE	16.091	2.866	18.957
SD	15.404	1.961	17.365
CO	4.116	615	4.731
NO	4.136	667	4.803
SU	9.935	1.657	11.592
TOTAL	49.682	7.766	57.448

Fonte: autor a partir de TSE

Em 2016, segundo dados oficiais do Tribunal Superior Eleitoral (TSE), do total de eleitos, 49.682 se registraram como homens e 7.766 como mulheres no TSE, resultando em 57.448 eleitos com a variável "sexo" válida. O teste χ^2 indica se podemos ou não rejeitar a hipótese nula de independência entre as variáveis. O resultado é de χ^2 = 122,703 e α = 0,000, portanto, um resultado altamente significativo, permitindo a rejeição da hipótese nula de independência entre as variáveis. Ou seja, as variações de sexo dos vereadores eleitos não são independentes das regiões. Agora, para identificar a magnitude do efeito, calcularemos o Cramer's V. Como a variável com menor número de categorias é sexo, com duas categorias, não haverá efeito do fator (k-1) na fórmula, pois multiplicaríamos o número de casos por um.

$$v = \sqrt{\frac{\chi^2}{N.(k-1)}} = \sqrt{\frac{122{,}703}{57.448\ x\ (2-1)}} = 0{,}046$$

Assim, temos que a associação entre as duas variações é de 0,046 ou 4,6% apenas. Ou seja, relação entre região sobre a número de homens ou mulheres eleitos é de apenas 4,6%. Embora o χ^2 seja alto e significativo, percebe-se que o efeito não é tão forte como se poderia pensar inicialmente. Se elevarmos o Cramer's V ao quadrado, teremos a proporção de variação que é explicada pela associação, ou seja, teremos a magnitude da determinação de uma variável sobre a outra. No caso, V^2 = 0,0021. Se multiplicarmos o valor por 100, teremos que apenas 0,21% da variação de proporção de sexo é explicada pela região, ou seja, uma explicação nula.

IMPORTANTE: A prova de χ^2 só permite aceitar ou rejeitar a hipótese nula,

mas no caso de rejeitá-la, não é possível saber em que medida as duas variáveis estão relacionadas. O coeficiente de contingência ou coeficiente de correlação, obtido através de tabelas de contingência, é a prova adequada a ser aplicada depois de rejeitada a hipótese nula pelo χ^2. O anexo ao capítulo apresenta os valores padronizados para identificar o limite crítico a partir do Intervalo de confiança e dos graus de liberdade nos testes de χ^2 para comparação de uma distribuição real com uma teórica e para os testes entre duas variáveis independentes. No anexo a este capítulo encontra-se uma tabela de valores padronizados de χ^2 para rejeição da hipótese nula em função dos graus de liberdade e do intervalo de confiança, e posterior realização dos testes de associação.

Os testes de diferença de médias como o χ^2 e testes de associação como Cramer's V produzem coeficientes capazes de indicar as relações entre todas as categorias de duas variáveis. Mas, muitas vezes o pesquisador necessita identificar relações entre pares de categorias de duas variáveis distintas, o que não é possível com os testes apresentados até aqui. Para um nível acima de detalhamento das associações em que se identifiquem relações entre pares de variáveis são indicados, entre outros, os testes que serão apresentados a seguir.

1.2 Coeficiente Delta (Δ) para diferenças de F_O e F_E

Um coeficiente Δ é uma medida que mostra a existência ou não de valores "sobrando" em determinados pares de categorias. Portanto, ele só deve ser aplicado em comparações entre duas variáveis, chamadas aqui de X e Y, mas poderiam ser quaisquer outras letras. Para deduzir se existe ou não alguma relação entre os pares de categorias das variáveis X e Y, comparam-se as frequências observadas a uma tabela com uma distribuição teórica na qual as distribuições são independentes – chamada de tabela de frequências esperadas. A hipótese estatística inicial é de que não há dependência de variações entre X e Y. Para tanto, é preciso que a frequência esperada seja a mesma ou esteja muito próxima da frequência observada, o que não nos permitirá rejeitar a hipótese nula. Por outro lado, a existência de diferenças significativas entre as

frequências observadas e as esperadas nos permite rejeitar a hipótese nula e passamos a considerar que as variações de X e Y apresentam alguma dependência. O passo seguinte é medir a força da relação ou dependência entre as duas variáveis.

O termo técnico para descrever a inexistência de relação entre duas variáveis é “independência estatística”. Portanto, temos independência estatística quando X e Y são estatisticamente independentes, o que ocorre quando as probabilidades das células esperadas igualam os produtos das probabilidades marginais relevantes. Nesse caso, também é indicado que as probabilidades de ocorrência de uma categoria em uma variável são as mesmas que as demais, independente da categoria da outra variável com a qual está ligada. Em outras palavras, a frequência de casos na categoria da segunda variável não faz diferença para a primeira para que a relação não tenha efeito estatístico. O que é diferente de dizer que não tem efeito algum.

O coeficiente Δ serve para indicar a existência de diferenças entre pares de frequências de uma tabela de contingência. Normalmente, é utilizado em tabelas quádruplas, para variáveis dicotômicas. No entanto, nada impede que também seja aplicado a cruzamentos entre variáveis com mais de duas categorias. O objetivo aqui é identificar se determinada característica conjunta de X e Y ocorre mais ou menos vezes do que seria esperado. Se isso acontecer, não podemos considerar as variáveis independentes para esse par de categorias. Aqui, são usadas as probabilidades observada e esperada nas comparações. Então, Δ pode ser representado pela seguinte fórmula:

$$\Delta = \text{Prob. Observada} - \text{Prob. Esperada}$$

Onde:

Prob. Observada = o valor da frequência de casos para determinado par de categorias;

Prob. Esperada = a multiplicação das marginais da tabela, dividido pelo número total de casos.

Digamos que o pesquisador queira encontrar o Δ para a probabilidade de homens que foram eleitos vereador com escolaridade superior em 2016. A hipótese é que os homens com escolaridade superior sejam em maior proporção que mulheres com

escolaridade superior, entre os eleitos. O primeiro passo é montar uma tabela de contingência entre sexo dos vereadores eleitos e escolaridade declarada no registro ao TSE em 2016, como a que segue (para facilitar o exercício, foram agregadas as categorias "analfabeto" e "lê e escreve" e todas as categorias "incompleto" e "completo" em categoria única do nível de escolaridade. Assim, "ensino fundamental completo" e "ensino fundamental incompleto" se transformaram em "ensino fundamental", e sucessivamente):

Tabela 1.2.1. Distribuição das proporções de vereadores eleitos por escolaridade e sexo em 2016

ESCOLARIDADE	MULHER	HOMEM	TOTAL
ANALFAB./LÊ E ESCREVE	62 (0,001)	960 (0,017)	1.022 (0,018)
ENSINO FUNDAMENTAL	905 (0,016)	14.617 (0,254)	15.522 (0,270)
ENSINO MÉDIO	2.757 (0,048)	20.775 (0,362)	23.532 (0,410)
SUPERIOR	4.042 (0,070)	13.330 (0,232)	17.372 (0,302)
TOTAL	7.766 (0,135)	49.682 (0,865)	57.448 (1,000)

Fonte: autor a partir de TSE

Olhando as marginais da tabela, é possível perceber que de maneira geral os homens representam quase nove em cada dez eleitos em 2016 (0,865), enquanto as mulheres ficam em uma proporção de apenas 0,135. Já as marginais das linhas mostram que a escolaridade dos eleitos também apresenta grande desproporção. A categoria "sem escolaridade" representa apenas 0,018 dos eleitos, passa a 0,270 a proporção de eleitos com ensino fundamental, sobe para 0,410 em ensino médio e cai para 0,302 para ensino superior. Se olharmos para o corpo da tabela, encontramos os valores das participações proporcionais de cada par de categorias. No caso, o que interessa aqui são os homens com escolaridade superior, que apresentam proporção de 0,232 (13.330 / 57.448 = 0,232). A questão é saber se essa proporção equivale a uma distribuição independente para o par de categorias ou se, ao contrário, ela indica a existência de algum grau de associação. O primeiro passo é encontrar a probabilidade esperada para eleitos com escolaridade superior. Como a tabela 1.2 já nos dá as proporções basta multiplicar a marginal da linha "escolaridade superior" pela marginal da coluna "homem": 0,302 x 0,865 = 0,261. Então, a proporção esperada de homens eleitos por partido de direita é de 0,261. Para conhecer o Δ, aplica-se a fórmula:

$$\Delta\text{pd} = \text{Prob. Observada} - \text{Prob. Esperada} = 0{,}232 - 0{,}261 = -0{,}029$$

O resultado do Δpd (delta para prefeitos de direita) é de -0,029, portanto, muito próximo de zero, ficando praticamente idêntico ao que seria uma distribuição independente. Não podemos dizer que foram eleitos mais homens com escolaridade superior do que mulheres com a mesma escolaridade, pois o valor observado não se distanciou do esperado, o que indica variações independentes entre as duas variáveis para essas categorias. A tabela 1.2.2 a seguir mostra os valores Δ para todas as categorias do exemplo.

Tabela 1.2.2. Valores de Δ para todos os pares de categorias da tabela 1.2

ESCOLARIDADE	MULHER	HOMEM
ANALFAB./LÊ E ESCREVE	-0,001	0,001
ENSINO FUNDAMENTAL	-0,021	0,021
ENSINO MÉDIO	-0,007	0,007
SUPERIOR	0,029	-0,029

Fonte: autor

A tabela acima mostra que as variações das proporções de homens e mulheres eleitos por escolaridade são muito pequenas, girando em torno do valor teórico. Todas ficam abaixo de 3% de diferença. Portanto, ainda não podemos dizer que existam diferenças significativas na eleição de homens ou mulheres entre os níveis de escolaridade, embora as categorias ensino superior e ensino fundamental apresentem as maiores diferença de Δ para homens e mulheres – com sinal invertido. Para ensino fundamental, o coeficiente é positivo para homens e negativo para mulheres, indicando que houve maior ocorrência de homens nessa categoria do que mulheres, enquanto que para escolaridade superior o sinal negativo é para homens.

Como a variável sexo possui apenas duas categorias (homem ou mulher), os coeficientes por grau de escolaridade serão os mesmos, havendo diferenças apenas entre os sinais positivo ou negativo. O sinal do Δ indica a direção. Se as diferenças proporcionais fossem maiores, poderíamos dizer que foram eleitas mais mulheres com escolaridade superior do que homens, proporcionalmente. No próximo tópico discutire-

mos como identificar os pontos críticos a partir dos quais podemos afirmar que não há independência de variações entre pares de categorias. Por enquanto, olhando para a tabela 1.2.2 acima, é possível concluir que entre escolaridade e sexo dos vereadores eleitos em 2016, os níveis "sem escolaridade" (analfabeto e lê e escreve) e "escolaridade média" apresentam Δ muito próximos a zero, indicando que proporcionalmente para esses dois níveis foram eleitos tantos homens quanto mulheres. Já para escolaridade média o Δ indica uma diferença em favor dos homens e na escolaridade superior a diferença é favorável às mulheres, em função da inversão dos sinais positivo e negativo entre os sexos dos eleitos.

O fato de não encontrarmos dependência de variação entre as categorias não é um problema tão grande, pois um Δ diferente de zero também não significaria muita coisa. Isso porque esse é um coeficiente muito rústico. Seu forte não é a precisão. Ele apresenta dois sérios limitadores para a interpretação estatística:

i) é sensível ao tamanho da amostra. Se dobrássemos o N no exemplo anterior o valor de cada Δ também seria o dobro. Isso impossibilita a comparação de coeficientes Δ em amostras com N diferentes, e;

ii) O coeficiente Δ não possui um limite superior. No limite inferior o valor é zero, mas não é possível saber até quanto se pode chegar ao outro limite, tanto positivo, quanto negativo. Isso impossibilita estabelecer magnitudes comparativas (Pestana & Gageiro, 2014).

Por outro lado, a vantagem do coeficiente Δ é a simplicidade do seu cálculo para tabelas de contingência. No próximo tópico, serão apresentados testes mais sofisticados que permitem maior detalhamento nas análises de relações entre pares de categorias de duas variáveis distintas.

1.3 Testes de associação entre categorias de variáveis nominais em tabelas de contingência (Resíduos Brutos e Resíduos Padronizados)

O coeficiente Δ é o primeiro obtido a partir da diferença entre o valor esperado e o que foi realmente observado na distribuição. Essa é a base para uma série de des-

crições e inferência estatísticas comumente usadas na Ciência Política. Alguns testes ganham em detalhamento. Como veremos mais adiante, o coeficiente Gama considera as direções das categorias ordinais para o estabelecimento das somas de valores de pares consistentes e de pares inconsistentes para identificação da existência ou não de associação entre as variáveis.

Porém, antes do Gama, quando estamos analisando as relações entre as categorias das variáveis nominais, não existe uma organização ordinal e transitiva entre as categorias. Logo, levar em conta a posição da categoria nas linhas ou nas colunas em relação a seus "vizinhos" não faz sentido nesse caso. Quando usamos o teste χ^2, temos apenas um coeficiente que indica a existência de relação entre pelo menos duas das categorias das variáveis testadas. Mas, pode ser que queiramos uma indicação mais precisa sobre quais das categorias das duas variáveis apresentam relações mais fortes, ou seja, quais contribuem de fato para a rejeição da hipótese de independência entre as variáveis. Para conhecer o peso das relações entre cada par de categorias, recomenda-se, entre outras, a análise de resíduos em tabelas de contingência.

Uma tabela de contingência é uma tabela que sumariza as frequências de ocorrências para cada par de categorias de duas variáveis X e Y quaisquer. O conceito de independência entre variáveis tem como princípio que a distribuição observada das frequências nas casas da tabela de contingência é muito próxima da distribuição esperada das frequências, ou seja, que a diferença entre as frequências observada e esperada deve estar próxima de zero no caso de distribuições independentes. Se houver diferenças entre a distribuição esperada e observada, podemos pensar em rejeitar a hipótese nula e considerar a existência de alguma associação entre pelo menos um par de categorias das duas variáveis testadas. Um exemplo de tabela de contingência para duas variáveis nominais é a que segue (tab. 1.3.1), entre ideologia do partido do eleito a vereador em 2016 por região do país. Como se trata apenas de um exemplo, não entrarei na discussão sobre que partidos são de direita, de centro e de esquerda no Brasil, pois isso com certeza daria uma série de outros livros. Apenas informo que os critérios utilizados são os mais aceitos pela literatura da Ciência Política brasileira, agrupando os partidos por ideologia a partir dos eixos participação do Estado na economia e defesa de valores morais.

Tabela 1.3.1. Distribuição do nº de vereadores eleitos por ideologia partidária e região do País

IDEOLOGIA	NE	SD	CO	NO	SU	TOTAL
ESQUERDA	4.518	2.622	730	865	2.488	11.223
CENTRO	3.983	5.247	1.529	1.187	4.172	16.118
DIREITA	10.546	9.595	2.485	2.767	4.983	30.376
TOTAL	19.047	17.464	4.744	4.819	11.643	57.717

Fonte: autor a partir do TSE

Aplicando a fórmula do χ^2, teríamos um coeficiente de χ^2 = 1.377,501 para as distribuições de frequências das duas variáveis. Considerando que temos 8 graus de liberdade [(3-1) x (5-1) = 8], se olharmos na tabela padronizada do χ^2 no anexo 1 deste capítulo, perceberemos que o limite crítico para o graus de liberdade e intervalo de confiança de 95% é de 15,50. Portanto o valor χ^2 = 1.377,501 fica muito acima desse limite. Isso nos permite rejeitar a hipótese nula e aceitar a possibilidade de que as variações das categorias das duas variáveis não são totalmente independentes. Nesse caso, podemos considerar que as variações de eleitos por ideologia e região do país apresentam alguma dependência, não sendo aleatórias as distribuições, pois o χ^2 aponta para uma possibilidade muito abaixo do limite crítico para a aceitação da aleatoriedade. No entanto, usando o χ^2 pararíamos as análises aqui. Este coeficiente não nos permite especular sobre porque essa dependência ocorre, por exemplo. Outra questão que fica sem resposta é se existe dependência entre todos os pares de categorias das variáveis ou em apenas parte deles. Por exemplo, é possível pensar que eleitos por partidos de esquerda concentrem-se mais em algumas regiões do País, enquanto eleitos por partidos de centro estão mais distribuídos em todas as regiões. O coeficiente χ^2 não permite verificar a validade dessas afirmações. É preciso complementar as análises a partir dos testes de resíduos em tabelas contingenciadas, também chamados de resíduos brutos.

1.3.1 Cálculo dos Resíduos Brutos (Rb)

Você já deve ter notado que os resíduos brutos nada mais são do que a diferença entre a Frequência Observada (Fo) e a Frequência Esperada (Fe). Eles ajudam

a evitar erros comuns na interpretação dos valores observados, pois quando as marginais não têm os mesmos valores, as frequências totais podem ser enganosas (Pereira, 2004). Por exemplo, na tabela anterior podemos olhar para a linha dos partidos de direita e comparar com a coluna da região sul concluindo que a direita elegeu mais que os partidos de centro nessa região (4.983 da direita no sul contra 4.172 do centro no sul). No entanto, se considerarmos as diferenças das marginais das linhas, perceberemos que proporcionalmente ao total de vereadores de direita e de centro, a participação dos partidos de centro no sul em relação ao total dos vereadores de centro foi maior do que a participação dos partidos de direita. São 4.172 de centro no sul de um total de 16.118, contra 4.983 de direita no sul de um total de 30.376. Por definição, o resíduo bruto de uma casa é a diferença entre a Fo e Fe. Na linguagem matemática seria:

$$R_b = F_o - F_e$$

Já a Frequência esperada é calculada da seguinte forma:

$$F_e = \frac{M.Linha\ x\ M.Coluna}{N}$$

Onde:

M. linha = marginal da linha;

M. coluna = marginal da coluna;

N = total de casos na tabela.

Para o exemplo anterior, a Fe do par eleito por “partido de esquerda” e estar na “região norte” seria:

$$F_e = \frac{M.Linha\ x\ M.Coluna}{N} = \frac{11.223\ x\ 4.819}{57.717} = 937{,}04$$

E o resíduo bruto para esse par de casos seria:

$$R_b = F_o - F_e = 865 - 937{,}04 = -72{,}04$$

A interpretação desse resultado é que na distribuição observada os partidos de esquerda tiveram cerca de 72 vereadores eleitos a menos na região norte do que deveria existir caso as distribuições fossem independentes.

Tabela 1.3.2. Valores de Freq. Esperada e Resíduos Brutos para ideologia do partido do vereador eleito e região do país em 2016

IDEOL	Frequência Esperada					Resíduo Bruto				
	NE	SD	CO	NO	SU	NE	SD	CO	NO	SU
ESQ.	3.703,67	3.395,85	922,46	937,05	2.263,97	814,33	-773,85	-192,46	-72,05	224,03
CEN.	5.319,05	4.876,98	1.324,81	1.345,75	3.251,41	**-1.336,05**	370,02	204,19	-158,75	**920,59**
DIR.	10.024,29	9.191,16	2.496,73	2.536,20	6.127,62	521,71	403,84	-11,73	230,8	-1.144,62

Fonte: autor

Agora, tendo gerado os resíduos brutos, já podemos identificar as principais diferenças nas concentrações dos valores. Olhando para a tabela como um todo, percebemos que os maiores resíduos brutos estão na região nordeste, onde partidos de esquerda elegeram 814,33 vereadores a mais do que o esperado e os partidos de direita elegeram 521,71 a mais. Já os partidos de centro elegeram menos 1.336,05 vereadores do que o esperado para o Nordeste, sendo este o maior valor negativo da tabela. A leitura também pode ser feita nas linhas. Por exemplo, os partidos de centro elegeram mais que o esperado no Sudeste, Centro-oeste e Sul, nesta última região é a maior diferença positiva da tabela (920,59 vereadores eleitos por partidos de centro a mais do esperado para a região sul). Seguindo a leitura por linhas, o melhor desempenho dos partidos de esquerda foi na região nordeste, com 814,33 eleitos a mais do que o esperado, assim como partidos de direita no nordeste também tiveram seu melhor desempenho, com 521,71 eleitos a mais que o esperado. Partidos de esquerda tiveram resultados positivos apenas no nordeste e no sul, enquanto partidos de direita tiveram bons desempenhos no nordeste, sudeste e norte.

A vantagem da análise dos resíduos brutos é que os valores estão na unidade de análise, no caso, em número de vereadores eleitos. Por outro lado, isso pode ser uma desvantagem, pois não permite comparações diretas com resíduos de outras dimensões ou até mesmo magnitudes. O problema dos resíduos brutos é serem pouco informativos, pois não apresentam variância constante. Em outras palavras, são não-padronizados. Também não permitem a verificação de pontos extremos (*outliers*) por não poderem ser comparados diretamente. Para resolver esse problema é preciso padronizar os resíduos. Com isso, torna-se possível verificar quem são as relações de casos mais extremos, quais são as maiores concentrações de casos e como se trata de valores adimensionais é possível comparar resíduos padronizados de variáveis de distintas dimensões ou magnitudes.

Até padronizarmos os resíduos não é possível saber quais são os resíduos com tamanho suficiente para serem considerados estatisticamente significativos e quais estão abaixo do limite crítico. O χ^2 da tabela de contingência mostrou que as variáveis apresentam dependência de variações. Os resíduos brutos indicaram diferenças que em alguns pares chegam a ser dez vezes maiores que em outros. Se por um lado os partidos de centro apresentaram o maior resíduo bruto (+920,59) na região sul, eles também ficaram com o maior resíduo negativo (-1.336,05) na região nordeste. Será que os dois podem ser considerados válidos no teste de associação entre as variáveis? Já o menor resíduo de toda a tabela, aquele que mais se aproximou de zero, foi para partidos de direita na região centro-oeste (-11,73). Até que ponto os resíduos podem ser usados como indicadores de diferenças significativas? Para responder a essas perguntas é preciso padronizar os valores dos resíduos e, a partir de um limite pré-estabelecido, indicar se os resíduos são significativos ou não.

Os Resíduos Padronizados (Rp) usam os valores equivalentes a *z-score* para permitir a identificação dos pares que estão acima do limite crítico e, portanto, apresentam "acúmulos" de frequências acima ou abaixo do que seria estatisticamente esperado significativo se a distribuição dos casos entre as variáveis fosse independente. Porém, antes de entrarmos nas explicações específicas do Resíduo Padronizado é preciso um lembrete: só faz sentido calcular o resíduo padronizado quando o resultado do χ^2 de uma tabela de contingência é significativo. Se o resultado não for estatisticamente significativo, todos os valores de Resíduos Padronizados ficarão abaixo do limite crítico, ou seja, não serão significativos. Por outro lado, se tivermos um χ^2 significativo em uma tabela de contingência, devemos calcular os resíduos padronizados para identificar quantos e quais pares de casos estão acima do limite crítico, quer dizer, concentram ou não mais casos do que o esperado se as variáveis fossem independentes.

1.3.2 Cálculo dos Resíduos Padronizados (Rp)

A análise de resíduos padronizados nada mais é do que a verificação dos valores que representam a relação biunívoca (nas duas direções) com probabilidade de chances de ocorrências. Ou seja, são os valores que sobram (para mais ou para menos) quando a

distribuição entre o valor observado e o esperado fica distante de zero, ou seja, a variação não é aleatória. Ao se estabelecer 95% de intervalo de confiança, essas chances de ocorrência são de ± 1,96, valor que serve de ponto de corte para o nível de significância de falta ou excesso de ocorrência entre as variáveis, o que permite distinguir os valores de pares casuais dos não casuais. Como o valor na tabela *z-score* para o intervalo de confiança de 95% é de 1,96, pode-se considerar que valores de resíduos padronizados acima de +1,96 ou abaixo de –1,96 apresentam excessos ou ausência de casos significativos, sendo, portanto, responsáveis pelas relações não aleatórias apontadas pelo coeficiente χ^2.

Todo resíduo, seja ele bruto ou padronizado, serve para indicar as diferenças entre o valor observado e o valor esperado em uma distribuição de frequências. O cálculo dos resíduos padronizados é bastante simples e quase intuitivo. Os valores são padronizações, ou seja, transformações adimensionais dos resíduos brutos. Um resíduo padronizado (Rp) é calculado a partir da padronização dos resíduos brutos para que passem a ter variância igual e apresentem-se de maneira adimensional, transformando-se em um coeficiente. Por ser padronizado, o Rp apresenta variância constante, o que permite a comparação direta entre valores de diferentes magnitudes. Se a análise é feita a partir de uma grande amostra (n > 120) e intervalo de confiança de 95% (z = 1,96), qualquer resíduo acima de 1,96 deve ser considerado estatisticamente significativo, ou seja, o resíduo encontrado naquela relação biunívoca é maior do que supunha a hipótese de independência entre as variações das duas variáveis. Podemos calcular os resíduos padronizados em tabelas de contingência de variáveis categóricas a partir da seguinte fórmula:

$$R_p = \frac{R_b}{\sqrt{F_e}}$$

Onde:

Rp = resíduo padronizado

Rb = resíduo bruto

Fe = frequência esperada

Usando o mesmo exemplo das distribuições de vereadores eleitos por ideologia partidária e região do País, temos que para a primeira célula da tabela de contingência –

L1, partido de esquerda, C1, região nordeste – o seguinte cálculo de Resíduo Padronizado:

$$R_{p(l1,c1)} = \frac{R_b}{\sqrt{F_e}} = \frac{814{,}33}{\sqrt{13.703{,}67}} = 6{,}95$$

Considerando o limite crítico de 1,96, podemos afirmar que os resíduos de partidos de esquerda na região nordeste são estatisticamente significativos. Ou seja, a esquerda elegeu mais (sinal positivo) vereadores no nordeste do que esperado se a distribuição dos eleitos de esquerda fosse aleatória por região do país. A tabela a seguir mostra todos os resultados de Resíduos padronizados para as duas variáveis.

Nela podemos perceber que um dos valores não é estatisticamente significativo. Ou seja, esse par de categorias não deve ser considerado com variações independentes, ainda que apresente resíduos brutos. É o caso dos vereadores eleitos por partidos de direita no Centro-Oeste (-0,235), bastante abaixo do limite de -1,96. Dos outros 14 pares de categorias que ficaram acima do limite crítico, oito apresentaram resíduos positivos e seis com resíduos negativos (indicados pelas cores azul e vermelha na tabela 1.3.3). Os resíduos padronizados mais intensos, como esperado, são os mesmos que os resíduos brutos, porém, agora com valores adimensionais. É o caso da eleição de vereadores de centro no Nordeste (-18,319) e partidos de centro no Sul (+16,145). Partidos de esquerda no Nordeste apresenta o segundo maior resíduo positivo (+13,381) e a direita na região Sul apresenta o segundo maior resíduo negativo (-14,622). Excetuando este último, os resíduos por região para partidos de direita são os que ficam mais próximos de zero, indicando que esta posição ideológica foi a que mais se aproximou de uma distribuição independente de região para todo País. Já os partidos de esquerda tenderam a concentrar eleitos no Nordeste, enquanto partidos de centro elegeram mais na região Sul.

Tabela 1.3.3. Resíduos Padronizados para ideologia do partido do eleito por região

IDEOLOGIA	NE	SD	CO	NO	SU
ESQUERDA	13,381	-13,28	-6,337	-2,354	4,708
CENTRO	**-18,319**	5,298	5,61	-4,327	**16,145**
DIREITA	5,211	4,212	-0,235	4,583	-14,622

Fonte: autor

Não vem ao caso, aqui, discutir quão contra intuitivo para o senso comum é a tabela 1.3.3, pois, de acordo com o senso comum, o eleitor nordestino tenderia a ser o mais refratário em relação à ideologia de esquerda em função das relações históricas com lideranças políticas ligadas a partidos que representam o centro ou a direita regional. A literatura especializada em eleições no Brasil tem demonstrando como nas últimas décadas o perfil de voto regional foi se transformando no Brasil, em especial após a chegada do principal partido de esquerda do País, o PT, ao governo federal (Cervi, 2016).

Como os resíduos são padronizados, também é possível fazer a leitura comparando os valores na mesma coluna, ou seja, entre as regiões. Nesse caso, teríamos que no Nordeste há um predomínio de vereadores de partidos de esquerda, no Sudeste há uma presença quase equilibrada entre partidos de centro e de direita, no Centro-Oeste predominam vereadores de partidos de centro, no Norte estão os vereadores de partidos de direta e no Sul voltam a eleger mais os partidos de centro. Em resumo, os resíduos padronizados foram necessários para a identificação individualizada da concentração de valores em pares de casos – acima ou abaixo – do esperado e dentro ou fora do limite crítico da significância estatística. Até então, o que tínhamos encontrado era um coeficiente que representasse o conjunto das relações entre todos os pares de casos.

1.3.3 Cálculo dos Resíduos Padronizados para análises temporais

Uma das principais limitações das técnicas quantitativas de análises temporais (as chamadas séries temporais) é a necessidade de um número mínimo de observações no tempo muito alto. Normalmente, é aceitável do ponto de vista estatístico pelo menos 120 pontos observados ao longo do tempo para uma análise consistente. Na maioria das vezes, os objetos de análise da ciência política não possuem todos esses pontos de observação no tempo. Isso é muito difícil em análises eleitorais, pois a distância entre as medições é grande, bianual, quadrienal ou até mais. Portanto, precisaríamos de dois séculos ou mais com dados disponíveis para podermos usar as técnicas tradicionais nesse caso. Mas, se a tabela de contingência for organizada em ordem temporal, os Resíduos Padronizados podem substituir as técnicas de séries

temporais com a vantagem de ser possível trabalhar com poucos pontos no tempo.

Quando comparados entre si, resíduos padronizados em uma tabela de contingência mostram as diferenças relativas entre cada par de categorias. Se uma das variáveis for temporal, a transição de uma categoria para outra indica uma mudança no tempo. Assim, diferenças de resíduos apontam para maior ou menor concentração de casos em determinado momento do tempo. Mas, atenção, os resíduos não são capazes de indicar quanto da mudança no tempo seguinte (t_1) é consequência ou "memória" da quantidade da característica no tempo anterior (t_0). Conhecer a proporção da característica que influencia o tempo seguinte apenas é possível usando as técnicas de análise de séries temporais que decompõem os valores – nesse caso, volta o problema da dependência do número mínimo de observações no tempo.

Usaremos para exemplificar o cálculo dos resíduos padronizados a distribuição dos deputados estaduais eleitos no Brasil entre 1998 e 2014. Como nesse período foram criados mais partidos dentro da posição de ideologia de centro e de direita, para não enviesar os resultados, vamos comparar os desempenhos apenas dos três maiores partidos que disputaram todas as eleições entre 1998 e 2014 (PT, PMDB e PSDB). Então, o objetivo da análise passa a ser verificar as mudanças nos resíduos padronizados de deputados estaduais eleitos entre as eleições. Nosso objetivo é saber se houve ou não variação estatisticamente significativa do desempenho de cada partido entre as eleições do período, usando a análise de resíduos. Para identificar os resíduos seguem-se os mesmos três passos do caso anterior, para cada um dos pares de resultados (Partido na linha x Ano na coluna). Para a primeira casa da tabela 1.3.4, número de deputados estaduais do PT eleitos por ano, teríamos o seguinte:

1º Passo (encontrar a Frequência Esperada)

$$\mathrm{F}_e = \frac{Ml\ x\ Mc}{N} = \frac{581\ x\ 332}{1.924} = 100{,}25$$

2º Passo (encontrar o Resíduo bruto)

$$R_b = F_o - F_e = 67 - 100{,}25 = -33{,}25$$

3º Passo (encontrar o Resíduo Padronizado)

$$R_p = \frac{Rb}{\sqrt{Fe}} = \frac{-33{,}25}{\sqrt{100{,}25}} = -3{,}321$$

Este mesmo procedimento deve ser repetido para todas as demais relações entre as categorias das variáveis analisada aqui. A princípio, olhando os valores brutos (N) na tabela 1.3.4 podemos perceber que os deputados estaduais eleitos pelo PT foram os que mais cresceram em número no período. Os do PMDB ficaram praticamente estáveis e os do PSDB apresentaram queda no período. Como as diferenças entre os totais por ano são relevantes – os três partidos elegem 332 deputados estaduais em 1998, esse número sobe para 433 em 2006, para depois cair a 341 em 2014 – seria equivocado comparar os valores de eleitos entre partidos ao longo do tempo. Os resíduos padronizados resolvem o problema das diferenças de totais. Na tabela 1.3.4, a linha Rp indica o resíduo para cada partido e ano de deputados estaduais eleitos, considerando apenas os três partidos. Ou seja, estamos interessados em saber se algum deles cresceu ou diminuiu em relação aos outros dois e não ao total geral.

Tabela 1.3.4. Número de deputados estaduais eleitos e resíduos padronizados para PT, PMDB e PSDB entre 1998 e 2014

PARTIDO	EST.	1998	2002	2006	2010	2014	TOTAL
PT	N	67	142	122	144	106	581
	Rp	-3,321	1,664	-0,766	1,844	0,298	
PMDB	N	133	130	161	147	139	710
	Rp	0,947	-1,704	0,096	-0,32	1,173	
PSDB	N	132	137	150	118	96	633
	Rp	2,179	0,21	0,632	-1,428	-1,528	
TOTAL		332	409	433	409	341	1924

Fonte: autor a partir do TSE

Como já sabemos, resultados acima de ± 1,96 para Resíduos Padronizados devem ser considerados estatisticamente significativos. Se positivo, significa que aquele par de categorias apresenta mais casos do que deveria caso as variáveis fossem independentes. Se negativo, ele concentra menos casos do que seria esperado. Na tabela acima, apenas dois resíduos mostram-se significativos, apesar das grandes variações de (N) dos partidos entre os anos. Em 1998, o resíduo padronizado do PT é -3,321, indicando que entre os três partidos ele foi o que elegeu menos parlamentares. O outro, em 1998, foi o resíduo para PSDB (+2,179), único resíduo positivo acima do limite crítico de toda a

série. Perceba que em termos de (N), em 1998 o PSDB elegeu um deputado estadual a menos que o PMDB. Ainda assim, o Rp do PSDB é maior que o do PMDB, isso porque os valores estão padronizados levando em conta as cinco eleições. De qualquer maneira, em geral os desempenhos dos três partidos, quando comparados entre si nas eleições para deputado estadual entre 1998 e 2014, não sofreram grandes diferenças, em especial nas disputas do final do período.

No entanto, é importante lembrar que os dados da tabela 1.3.4 estão agregados para todo o País. Se quisers, podemos desagregar por região para tornar as informações ainda mais detalhadas e sermos capazes de identificar se, quando comparados entre si, um dos três partidos ganhou ou perdeu mais deputados estaduais por região e ano eleitoral. A tabela 1.3.5 é resultado dos mesmos procedimentos já demonstrados acima, por isso estão apenas os resultados de N e Rp por região e ano.

Como se pode perceber a partir das indicações e cores para os resíduos estatisticamente significativos, na maior parte das eleições não houve diferenças estatisticamente significativas entre os números de deputados estaduais eleitos pelos três partidos. O PMDB oscilou, abaixo do limite crítico, sem uma tendência clara, em todas as regiões nas cinco disputas. O PT apresentou tendência de crescimento na região Norte e uma tendência constante de queda nos resíduos da região Sul durante o período, porém, todos os coeficientes ficaram abaixo do limite crítico nessas duas regiões. No Centro-Oeste, ele oscilou sem tendência clara. No Nordeste, ele passou de coeficiente negativo e estatisticamente significativo em 1998 (-2,466) para resíduo positivo, porém, abaixo do limite crítico, em 2014 (+0,795). E no Sudeste apresentou redução de resíduo negativo e estatisticamente significativo em 1998 (-2,327) para -0,217 em 2014. O PSDB, quando comparado aos outros dois partidos, oscilou sem tendência clara no Norte e Centro-Oeste entre 1998 e 2014. No Nordeste, ele apresentou queda comparativa, passando do resíduo positivo e não significativo de +1,161 em 1998, para -2,128 em 2014. No Sudeste, ele também apresenta queda comparativa no número de deputados estaduais eleitos, passando de resíduo positivo e significativo (+2,229) em 1998 para negativo e não significativo em 2014 (-0,656). Na região Sul, o PSDB apresenta oscilação abaixo do limite crítico, porém, com tendência de crescimento, passando de um resíduo negativo -0,298 no início da série temporal, para +0,421 no final do período de análise.

Tabela 1.3.5. Dep. estaduais eleitos e resíduos padronizados para PT, PMDB e PSDB por região do País entre 1998 e 2014

PARTIDO	EST.	NORTE 1998	2002	2006	2010	2014	TOTAL
PT	N	12	19	20	20	16	87
	Rp	-0,437	-0,054	0,039	0,386	0,005	
PMDB	N	21	25	26	26	24	122
	Rp	0,438	-0,38	-0,342	0,054	0,336	
PSDB	N	13	21	21	16	14	85
	Rp	-0,082	0,509	0,37	-0,455	-0,408	
TOTAL		46	65	67	62	54	294
PARTIDO	EST.	NORDESTE 1998	2002	2006	2010	2014	TOTAL
PT	N	16	33	34	43	28	154
	Rp	**-2,466**	-0,041	-0,244	**1,975**	0,795	
PMDB	N	47	40	44	41	41	213
	Rp	1,003	-0,881	-0,719	-0,46	1,329	
PSDB	N	43	47	50	31	18	189
	Rp	1,161	0,972	0,984	-1,294	**-2,128**	
TOTAL		106	120	128	115	87	556
PARTIDO	EST.	CENTRO-OESTE 1998	2002	2006	2010	2014	TOTAL
PT	N	3	9	9	9	9	39
	Rp	-1,532	0,069	0,203	0,489	0,724	
PMDB	N	17	14	21	19	14	85
	Rp	0,403	-1,18	0,623	0,572	-0,361	
PSDB	N	17	23	14	12	14	80
	Rp	0,654	1,168	-0,784	-0,931	-0,134	
TOTAL		37	46	44	40	37	204
PARTIDO	EST.	SUDESTE 1998	2002	2006	2010	2014	TOTAL
PT	N	19	50	37	45	34	185
	Rp	**-2,327**	1,877	-0,487	0,951	-0,217	
PMDB	N	26	27	33	30	34	150
	Rp	-0,022	-0,737	0,088	-0,296	1,007	
PSDB	N	49	35	47	39	35	205
	Rp	**2,229**	-1,153	0,388	-0,65	-0,656	
TOTAL		94	112	117	114	103	540
PARTIDO	EST.	SUL 1998	2002	2006	2010	2014	TOTAL
PT	N	17	31	22	27	19	116
	Rp	-0,054	1,619	-0,974	-0,08	-0,455	
PMDB	N	22	24	37	31	26	140
	Rp	0,266	-0,756	0,758	-0,363	0,108	
PSDB	N	10	11	18	20	15	74
	Rp	-0,298	-0,988	0,176	0,6	0,421	
TOTAL		49	66	77	78	60	330

Fonte: autor a partir do TSE

Como é possível perceber nos dados da tabela acima, quando separamos os desempenhos por regiões ganhamos em detalhamento de informações, porém, torna-se mais complexa a explicação das relações e, portanto, mais fácil de tornar os resultados

incompreensíveis. Em resumo, se no final do período o desempenho comparativo dos três partidos para deputado estadual apresentou resíduos muito próximos entre si, em 1998 o PSDB apresentava melhor desempenho do que o PT e o que explica essa diferença é que no final da década de 1990 o PT elegia menos deputados no nordeste e no Sudeste, enquanto o PSDB elegia mais no Sudeste. Em 2014, o PSDB elegeu menos no Nordeste, equiparando as diferenças iniciais com o PT.

O objetivo deste capítulo foi apresentar algumas técnicas simples para cálculo de coeficientes específicos aplicados a variáveis categóricas e para dados secundários, a partir de tabelas de contingência. Essas ferramentas são úteis para o pesquisador que pretende trabalhar com informações extraídas de relatórios ou publicações sobre os quais não é possível acessar o banco de dados primário. A recomendação geral é que o pesquisador use diferentes técnicas para complementar as informações obtidas isoladamente. Dependendo do tipo de variável, faça um teste de χ^2 de início e, se os coeficientes forem estatisticamente significativos, agregue uma análise de Resíduos Padronizados. Com isso, você poderá tirar conclusões não apenas sobre as variações das variáveis, mas também para os pares de relações entre as categorias. No próximo capítulo, discutiremos um teste estatístico específico para verificar a associação entre duas variáveis binárias, ou seja, para testar a força da relação entre as categorias em uma tabela quádrupla (2x2), que é simples, rápido e fornece coeficientes bastante explicativos. Quando aplicados a tabelas quádruplas, os testes dispensam a análise de resíduos individuais. Por fim, vale (re)lembrar que o conjunto de testes apresentado aqui abrange uma pequena parte das ferramentas estatísticas disponíveis para análise de dados categóricos.

1.4 Referências bibliográficas do Capítulo I

Cervi, E. U. (2016). *PSDB & PT em eleições nacionais*. Salamanca/Curitiba: Flacso-es/CPOP.

Pereira, J. C. R. (2004). *Análise de Dados Qualitativos*. São Paulo: EdUSP.

Pestana, M. H., & Gageiro, J. N. (2014). *Análise de Dados Para Ciências Sociais*. Lisboa: Ed. Sílabo.

1.5 Exercícios propostos do Capítulo I

Considere a tabela de contingência a seguir para o cruzamento entre as variáveis "Sexo do eleito" e "Partido" para o número de deputados estaduais eleitos em 2014. Atenção, trata-se de um exercício, portanto, faça os três cálculos indicados abaixo, independente do nível de significância da independência das variações:

SEXO	PARTIDO			TOTAL
	PT	PMDB	PSDB	
HOMEM	79	686	581	1.346
MULHER	21	238	244	503
TOTAL	100	924	825	1.849
χ^2 = 5,263 (0,072)				

1.5.1 O Coeficiente V de Cramer;

1.5.2 Os valores de Delta para todos os pares;

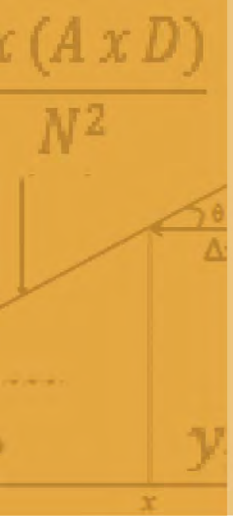
1.5.3 Os resíduos padronizados para todos os pares.

Interprete os resultados considerando a pergunta: É possível dizer que existiram diferenças estatisticamente significativas entre as proporções de mulheres eleitas por PT, PMDB e PSDB para deputadas estaduais em 2014?

ANEXO DO CAPÍTULO I

ANEXO 1.1 – VALORES PADRONIZADOS DA DISTRIBUIÇÃO DO χ^2

GL	0,995	0,975	0,9	0,5	0,1	0,05	0,025	0,01	0,005	0,001
1	0,000	0,001	0,016	0,455	2,706	3,841	5,024	6,635	7,879	10,827
2	0,010	0,051	0,211	1,386	4,605	5,991	7,378	9,210	10,597	13,815
3	0,072	0,216	0,584	2,366	6,251	7,815	9,348	11,345	12,838	16,266
4	0,207	0,484	1,064	3,357	7,779	9,488	11,143	13,277	14,860	18,466
5	0,412	0,831	1,610	4,351	9,236	11,070	12,832	15,086	16,750	20,515
6	0,676	1,237	2,204	5,348	10,645	12,592	14,449	16,812	18,548	22,457
7	0,989	1,690	2,833	6,346	12,017	14,067	16,013	18,475	20,278	24,321
8	1,344	2,180	3,490	7,344	13,362	15,507	17,535	20,090	21,955	26,124
9	1,735	2,700	4,168	8,343	14,684	16,919	19,023	21,666	23,589	27,877
10	2,156	3,247	4,865	9,342	15,987	18,307	20,483	23,209	25,188	29,588
11	2,603	3,816	5,578	10,341	17,275	19,675	21,920	24,725	26,757	31,264
12	3,074	4,404	6,304	11,340	18,549	21,026	23,337	26,217	28,300	32,909
13	3,565	5,009	7,041	12,340	19,812	22,362	24,736	27,688	29,819	34,527
14	4,075	5,629	7,790	13,339	21,064	23,685	26,119	29,141	31,319	36,124
15	4,601	6,262	8,547	14,339	22,307	24,996	27,488	30,578	32,801	37,698
16	5,142	6,908	9,312	15,338	23,542	26,296	28,845	32,000	34,267	39,252
17	5,697	7,564	10,085	16,338	24,769	27,587	30,191	33,409	35,718	40,791
18	6,265	8,231	10,865	17,338	25,989	28,869	31,526	34,805	37,156	42,312
19	6,844	8,907	11,651	18,338	27,204	30,144	32,852	36,191	38,582	43,819
20	7,434	9,591	12,443	19,337	28,412	31,410	34,170	37,566	39,997	45,314
21	8,034	10,283	13,240	20,337	29,615	32,671	35,479	38,932	41,401	46,796
22	8,643	10,982	14,041	21,337	30,813	33,924	36,781	40,289	42,796	48,268
23	9,260	11,689	14,848	22,337	32,007	35,172	38,076	41,638	44,181	49,728
24	9,886	12,401	15,659	23,337	33,196	36,415	39,364	42,980	45,558	51,179
25	10,520	13,120	16,473	24,337	34,382	37,652	40,646	44,314	46,928	52,619
26	11,160	13,844	17,292	25,336	35,563	38,885	41,923	45,642	48,290	54,051
27	11,808	14,573	18,114	26,336	36,741	40,113	43,195	46,963	49,645	55,475
28	12,461	15,308	18,939	27,336	37,916	41,337	44,461	48,278	50,994	56,892
29	13,121	16,047	19,768	28,336	39,087	42,557	45,722	49,588	52,335	58,301
30	13,787	16,791	20,599	29,336	40,256	43,773	46,979	50,892	53,672	59,702
31	14,458	17,539	21,434	30,336	41,422	44,985	48,232	52,191	55,002	61,098
32	15,134	18,291	22,271	31,336	42,585	46,194	49,480	53,486	56,328	62,487
33	15,815	19,047	23,110	32,336	43,745	47,400	50,725	54,775	57,648	63,869
34	16,501	19,806	23,952	33,336	44,903	48,602	51,966	56,061	58,964	65,247
35	17,192	20,569	24,797	34,336	46,059	49,802	53,203	57,342	60,275	66,619
36	17,887	21,336	25,643	35,336	47,212	50,998	54,437	58,619	61,581	67,985
37	18,586	22,106	26,492	36,336	48,363	52,192	55,668	59,893	62,883	69,348
38	19,289	22,878	27,343	37,335	49,513	53,384	56,895	61,162	64,181	70,704
39	19,996	23,654	28,196	38,335	50,660	54,572	58,120	62,428	65,475	72,055
40	20,707	24,433	29,051	39,335	51,805	55,758	59,342	63,691	66,766	73,403
41	21,421	25,215	29,907	40,335	52,949	56,942	60,561	64,950	68,053	74,744
42	22,138	25,999	30,765	41,335	54,090	58,124	61,777	66,206	69,336	76,084
43	22,860	26,785	31,625	42,335	55,230	59,304	62,990	67,459	70,616	77,418
44	23,584	27,575	32,487	43,335	56,369	60,481	64,201	68,710	71,892	78,749
45	24,311	28,366	33,350	44,335	57,505	61,656	65,410	69,957	73,166	80,078
46	25,041	29,160	34,215	45,335	58,641	62,830	66,616	71,201	74,437	81,400
47	25,775	29,956	35,081	46,335	59,774	64,001	67,821	72,443	75,704	82,720
48	26,511	30,754	35,949	47,335	60,907	65,171	69,023	73,683	76,969	84,037
49	27,249	31,555	36,818	48,335	62,038	66,339	70,222	74,919	78,231	85,350
50	27,991	32,357	37,689	49,335	63,167	67,505	71,420	76,154	79,490	86,660
51	28,735	33,162	38,560	50,335	64,295	68,669	72,616	77,386	80,746	87,967
52	29,481	33,968	39,433	51,335	65,422	69,832	73,810	78,616	82,001	89,272
53	30,230	34,776	40,308	52,335	66,548	70,993	75,002	79,843	83,253	90,573
54	30,981	35,586	41,183	53,335	67,673	72,153	76,192	81,069	84,502	91,871
55	31,735	36,398	42,060	54,335	68,796	73,311	77,380	82,292	85,749	93,167
56	32,491	37,212	42,937	55,335	69,919	74,468	78,567	83,514	86,994	94,462
57	33,248	38,027	43,816	56,335	71,040	75,624	79,752	84,733	88,237	95,750
58	34,008	38,844	44,696	57,335	72,160	76,778	80,936	85,950	89,477	97,038
59	34,770	39,662	45,577	58,335	73,279	77,930	82,117	87,166	90,715	98,324
60	35,534	40,482	46,459	59,335	74,397	79,082	83,298	88,379	91,952	99,608



Capítulo II

Teste de associação para tabelas quádruplas e para variáveis ordinais

A forma de organização inicial dos dados é determinante para o processo de análise e para os resultados que virão a seguir.

A análise de relações entre variáveis a partir de tabelas quádruplas (2x2) é uma excelente forma de realizar uma primeira aproximação das associações que o pesquisador espera encontrar no mundo empírico. Tabelas 2x2 sumarizam informações de um mundo bastante complexo. A principal contribuição para análises desse tipo de variáveis foi feita pelo estatístico inglês George Udny Yule quando, em 1911, publicou a primeira edição de "*An introduction to the theory of Statistics*". Nesse livro, foi demonstrado pela primeira vez um teste de associação entre variáveis binárias que ficou conhecido como Q de Yule (Q_{xy}). Devido à importância do teste, o texto foi republicado dezenas de vezes em diferentes línguas nas décadas posteriores à publicação da primeira edição. Neste capítulo, aprenderemos a calcular o Q_{xy} para cruzamentos entre duas e três variáveis binárias. Ao final, é apresentada a forma de calcular o coeficiente Gama, indicado para cruzamentos entre variáveis ordinais, com três ou mais categorias cada uma.

2.1 Teste Q de Yule (Q_{XY})

Como já apresentado no volume I do manual, uma variável binária ou dicotômica é aquela que possui apenas duas categorias, que representam a presença ou a ausência de determinada característica. Normalmente, a representação numérica das categorias é feita por 0 = ausência e 1 = presença. Pode ser aplicado, por exemplo, à variável Sexo, quando se quer testar determinada característica das mulheres, então: 1 = mulher e 0 = homem. Ou quando se quer dividir o total de eleitores em dois grupos, sendo: 1 = eleitores que votaram no candidato K na última eleição ou 0 = eleitores que não votaram no candidato K na última eleição. Até aqui identificamos duas variáveis dicotômicas: sexo e voto em determinado candidato. Digamos que nosso objetivo seja saber se o candidato K teve mais votos entre as mulheres quando comparado aos demais concorrentes. Nesse caso, precisaríamos cruzar as duas informações para ter quatro condições possíveis: a) é mulher e não votou em K; b) é mulher e votou em K; c) não é mulher e não votou em K e d) não é mulher e votou em K. Como existem quatro possibilidades em um cruzamento de duas variáveis dicotômicas, elas são organizadas em tabelas quádruplas (2x2). O passo seguinte é tentar identificar se a presença de terminada característica está associada à presença de característica em outra variável. No nosso exemplo, poderíamos nos perguntar se o fato de ser mulher está associado ou não a votar no candidato K. Um teste estatístico para medir a existência ou não de relação entre duas variáveis dicotômicas e, no caso de existir relação, a força e a direção da mesma foi proposto pelo estatístico inglês George Unde Yule em 1911. Conhecido por Q de Yule, é representado pela letra Q_{xy}, como veremos a seguir.

O teste de independência Q_{xy} serve para identificar se:

i) duas variáveis dicotômicas estão relacionadas entre si;

ii) de quanto é a intensidade da relação; e,

iii) se os resultados podem ser usados em generalizações para toda a população quando se está testando a associação em amostras.

Como é aplicado em tabelas quádruplas (com duas variáveis dicotômicas) e qualquer variável pode ser dicotomizada, trata-se de um coeficiente bastante útil e que pode ser obtido com a aplicação de fórmulas simples, dispensando o uso de programas

de computador. Uma variável pode ser dicotomizada quando se decide separar em dois grupos as categorias internas dela. Por exemplo, pode-se ter uma variável categórica na forma de Escala de Likert para avaliação de governo: Muito Boa, Boa, Regular, Ruim e Péssima. A dicotomização se dá quando o pesquisador divide os resultados entre Avaliação Positiva e as demais. Então, teríamos: 1 = (Muito Boa + Boa) e 0 = (Regular + Ruim + Péssimo), por exemplo. A dicotomização também pode ser a partir de uma variável escalar discreta, como idade em anos completos. Nesse caso, a opção pode ser usar o valor da mediana para dividir em dois grupos de igual tamanho. Então, se quiséssemos testar o efeito entre os mais velhos, teríamos: 0 = grupo dos mais novos, até a mediana e 1 = grupo dos mais velhos, a partir da mediana.

Também é possível dicotomizar distribuições de frequências a partir de dados secundários, como, por exemplo, usando informações de uma tabela de distribuição das intenções de voto a seis candidatos em uma eleição qualquer. Nesse caso, separa-se a frequência de respondentes que dizem votar em um candidato (representado pela letra K) e essa será a característica analisada (1). A soma de todas as demais receberá código zero. Ao final teremos apenas dois resultados possíveis: vota no candidato K ou não vota no candidato K.

O importante aqui é entender que qualquer variável pode ser dicotomizada desde que o processo seja defensável estatisticamente. Quando se tem duas variáveis dicotômicas, tais como votar ou não no candidato A e idade dos respondentes (jovem e não jovem) é possível aplicar os cálculos do coeficiente de Q_{xy} para identificar se as duas variáveis apresentam independência de variações ou se as variações delas estão associadas entre si. Se não forem independentes, significa que há alguma associação entre as características medidas. Então, o coeficiente também nos fornece a informação sobre o grau de associação entre elas, ou seja, a força. Uma terceira característica é a direção da associação. Quando as variações estão no mesmo sentido, ambas as variações passando de zero para um, por exemplo, o sinal é positivo. Quando existe associação, mas ela é cruzada, então o sinal será negativo. Por fim, o teste também mostra se os resultados obtidos em uma amostra são consistentes o suficiente para permitir a extrapolação para toda a população.

O mais comum quando se agregam variáveis escalares, proporcionais, ordinais

ou de intervalo é considerar X e Y o conjunto de valores Altos ou a presença da característica a ser medida e não-X e não-Y os valores Baixos ou a ausência da característica a ser medida. Essa convenção é importante em função do sinal do coeficiente de associação no resultado do teste. Uma inversão das posições significaria inverter um sinal de relação na mesma direção (positivo) por relação em direções opostas (negativo). As tabelas quádruplas são compostas por quatro células de frequências, quatro células com frequências marginais e uma célula de total, chamada de N. Cada uma das células de frequências recebe uma letra como nome, sendo A, B, C e D, como no quadro a seguir:

Quadro 2.1. Distribuição Quádrupla para cálculo do Q_{xy}

	Não-Y	**Y**	**Total**
X	**A**	**B**	Marginal X
Não-X	**C**	**D**	Marginal Não-X
Total	Marginal Não-Y	Marginal Y	Total de Casos (N)

Devem fazer parte das células de frequências apenas os casos válidos, o que sempre precisa ser explicitado aos leitores. As variáveis analisadas são chamadas de X e Y. As categorias de grupamento dicotômico das variáveis são chamadas, por consequência, de X e não-X; Y e não-Y. Em um exemplo de pesquisa sobre intenção de voto relacionada a sexo dos eleitores para saber se determinado candidato (K) recebe votos de mulheres, os respondentes que dizem votar no candidato K compõem as casas da linha X e aqueles que dizem votar em qualquer outro candidato fazem parte da linha Não-X. Já as eleitoras são Y e os eleitores são não-Y. As somas dos casos nas linhas (horizontais) e nas colunas (verticais) formam o que se chama de marginais. A somatória das marginais leva ao número total de casos analisados, representado pela letra N. Assim, teremos ao final uma tabela quádrupla que relaciona eleitores e não eleitores do candidato K com o fato de ser ou não ser mulher. O resultado apresentará se o candidato K tem uma concentração maior de votos entre as mulheres ou não.

Como todos os demais testes estatísticos probabilísticos, o Q_{xy} parte da hipótese inicial (H_0) de independência entre as variáveis. O que queremos identificar é se existe uma chance estatística forte suficiente para garantir baixas possibilidades de erro caso a hipótese nula (H_0) seja rejeitada e passemos a defender que existe alguma

relação entre as duas variáveis. No caso do exemplo, afirmar que o candidato K tem mais votos entre mulheres do que entre homens seria uma hipótese inicial de trabalho. Partiríamos do princípio de que não há diferença de sexo entre os eleitores do candidato K, ou seja, as duas variáveis são independentes, como prediz a H_0. Nosso objetivo é realizar os testes para verificar se temos condições suficientes de afirmar que há uma associação entre as duas variáveis – ser mulher e votar em K. Nesse caso, rejeitaríamos H_0 e assumiríamos que há uma probabilidade de que as duas variáveis estejam associadas, quer dizer, assumimos H_1. No próximo tópico, veremos como fazer isso para duas variáveis dicotômicas.

2.1.1 Teste de Independência Q de Yule (Q_{XY})

Os testes de independência visam identificar se as variações entre categorias de duas variáveis se dão de forma independente ou se elas aguardam alguma dependência entre si. A partir disso, se for identificada alguma dependência entre variações é possível pensar na existência de associação estatística. Se não, diz-se que a associação é nula, ou seja, as variáveis são independentes. Se sim, a associação pode ter diferentes intensidades: fraca, média, forte. Aqui, o teste de independência visa identificar a inexistência de relação das variações entre duas variáveis. Portanto, relembrando, a hipótese inicial é de independência. Se houver alguma relação entre as variações, então, nega-se a hipótese de independência e mede-se o grau de relação entre elas.

Nas tabelas quádruplas cada casa representa a frequência encontrada para um par de características (par Não-Y, X; par Não-Y, Não-X; par Y, X; par Y, Não-X). Se as variáveis forem independentes, a proporção de casos em cada par em relação ao total será a mesma ou muito próxima entre si, portanto, impedindo qualquer afirmação de associação entre as variáveis. Já se houver uma distorção razoável entre a frequência relativa de casos em um ou alguns pares em relação aos demais, podemos negar a independência e medir o grau de associação entre as categorias das variáveis. Então, o coeficiente Q_{xy} nos fornece duas informações importantes:

i) sobre a magnitude da relação, medida pelo tamanho do coeficiente. Quanto

mais próximo de ±1 mais forte será a associação; e

ii) a respeito da direção da relação. Se o sinal do coeficiente for positivo, então as duas categorias estão associadas e variam na mesma direção. Se o sinal for negativo, existe associação, mas as variações são em direções opostas.

O quadro a seguir representa os sinais predominantes nas associações Positivas e Negativas entre duas variáveis dicotômicas.

Quadro 2.2. Relação dos sinais nas tabelas quádruplas

Positiva

	Não-Y	Y
X	-	+
Não-X	+	-

Negativa

	Não-Y	Y
X	+	-
Não-X	-	+

No quadro acima, a associação positiva indica uma concentração de casos com a característica da variável X e com a característica da variável Y, mostrando que as presenças das características em X e Y "caminham na mesma direção". Já na associação negativa, a presença da característica na variável Y apresenta maior concentração de frequências na casa da ausência da característica na variável X, nesse caso, elas "caminham em direções opostas". Atenção para a diferença no uso dos termos "tende a ser" e "a maioria é". Nas análises probabilísticas deve-se fazer, sempre, a primeira afirmação ao invés da segunda.

O coeficiente Q_{xy} apresenta as características desejadas em um coeficiente de associação que pretenda medir a força e a direção da relação. Ele é insensível ao tamanho da amostra, portanto, seus resultados não oscilam em função do N da tabela quádrupla e ele apresenta limites superior e inferior pré-estabelecidos. Dessas duas características, são originados os seguintes postulados para o coeficiente de associação Q_{xy}:

a) O coeficiente deve ser igual a zero quando X e Y forem independentes; e

b) O coeficiente deve ser de no máximo + 1,00 para associação positiva e – 1,00 para associação negativa.

A partir desses postulados, Davis (1976) organizou as possíveis distribuições

de valores de Q_{xy} por grau de intensidade e forma adequada de interpretação, como segue proposto no quadro abaixo:

Quadro 2.3. Intervalos de valores para coeficiente Q_{xy}

Valor de Q_{xy}	Leitura adequada
+0,70 ou mais	Associação positiva muito forte
+0,50 a +0,69	Associação positiva forte
+0,30 a +0,49	Associação positiva moderada
+0,10 a +0,29	Associação positiva baixa
+0,01 a +0,09	Associação positiva desprezível
0	Nenhuma associação
-0,01 a -0,09	Associação negativa desprezível
-0,10 a -0,29	Associação negativa baixa
-0,30 a -0,49	Associação negativa moderada
-0,50 a -0,69	Associação negativa forte
-0,70 ou mais	Associação negativa muito forte

Fonte: Davis, 1976.

Quando o estatístico inglês G. Udny Yule apresentou uma proposta de coeficiente de correlação no início do século XX, ele pensou no teste respeitando as regras de distribuição dos casos apresentada no quadro 2.1 para aplicação aos resultados de uma tabela quádrupla. A primeira publicação do coeficiente foi em 1911 e Udny Yule o batizou de Q_{xy} em homenagem ao pioneiro astrônomo e estatístico belga Lambert Adolphe Jaques Quetelet (1796-1874). Com o tempo, o coeficiente passou a ser chamado de Q de Yule e sua fórmula é a seguinte (Yule & Kendall, 1937):

$$Q_{xy} = \frac{(BxC) - (AxD)}{(BxC) + (AxD)}$$

Onde:

A = é X e Não-Y;

B = é X e Y;

C = é Não-X e Não-Y;

D = é Não-X e Y.

Trata-se da divisão entre os produtos cruzados de uma tabela quádrupla, seguindo as posições das letras nas casas e das presenças e ausências dos casos nas linhas e colunas, como indicado no quadro 2.1, no início do capítulo.

Para exemplificar, vamos aplicar a fórmula para uma relação entre idade e sexo para todos os candidatos que disputaram as eleições nacionais de 2014 no Brasil, a partir das informações disponíveis no TSE. Em primeiro lugar, vamos dicotomizar a variável idade, usando a mediana como ponto de corte. No caso das eleições de 2014, os mais de 19 mil candidatos a todos os cargos em disputa apresentaram uma mediana de 47 anos, então, este será o valor que vai dividir a variável em dois grupos (Não-Y = abaixo da mediana e Y = acima da mediana). Queremos identificar a associação entre idade e sexo dos candidatos. No nosso caso, X = homem e Não-X = mulher. Vamos testar se existe independência entre as seguintes características: sexo do candidato e estar acima da mediana de idade. Nossa hipótese é que como os homens costumam apresentar maior permanência na política, eles tenderão a se concentrar mais no grupo dos candidatos mais velhos, acima da mediana de idade, enquanto as mulheres tenderão a se concentrar abaixo da mediana.

Se encontrarmos independência entre as características significa que os candidatos mais velhos distribuem-se igualmente entre homens e mulheres. A associação entre idade e sexo não é apenas ilustrativa aqui. A literatura em ciência política já tem demonstrado que candidatos mais experientes tendem a entrar nas campanhas com maiores chances de vitória. No caso das mulheres, a maior rotatividade de candidatas entre eleições poderia contribuir para uma concentração do sexo feminino entre os candidatos mais jovens. Aqui usaremos a idade como *proxy* de experiência. As frequências de casos das duas variáveis são apresentadas na tabela quádrupla a seguir:

Tabela 2.1. Sexo e idade de candidatos às eleições nacionais de 2014

	Abaixo mediana (Não Y)	Acima mediana (Y)	Total
Homem (X)	6.991	6.924	13.915
Mulher (Não X)	2.894	2.354	5.248
Total	9.885	9.278	19.163

Fonte: autor com dados do TSE.

Aplicando a fórmula do Q_{xy} teríamos que:

$$Q_{xy} = \frac{(B\ x\ C) - (A\ x\ D)}{(B\ x\ C) + (A\ x\ D)} = \frac{(6924\ x\ 2894) - (6991\ x\ 2354)}{(6924\ x\ 2894) + (6991\ x\ 2354)} = \frac{3581242}{36494870} = \mathbf{0,098}$$

Resposta: há uma associação da ordem de +0,098, ou +9,8%, entre ser homem e estar acima da mediada de idade. Como o coeficiente é positivo, isso indica que os deputados acima da mediana de idade tendem a estar acima da distribuição normal para o grupo dos deputados abaixo da mediana da idade. No entanto, ao olharmos o quadro de interpretação da magnitude do coeficiente (quadro 2.3) podemos dizer que existe uma associação positiva desprezível, pois fica abaixo de +0,10.

Uma das principais características do Q_{xy} é que, por ser o resultado de produtos cruzados de tabelas quádruplas, quando os produtos dos pares consistentes e inconsistentes estão relacionados, o Q_{xy} cresce. Além disso, o coeficiente tem limite superior em +1,00 e inferior em -1,00. No entanto, são necessários alguns cuidados na interpretação dos resultados. Por se tratar da divisão de produtos cruzados, quando uma das células for zero, o valor de Q_{xy} também será nulo, embora o cálculo matemático gere como resultado -1,00. Isso não significa necessariamente a existência de relação perfeita negativa. Aqui também vale a regra geral da estatística inferencial de que para tabelas de contingência é preciso ter uma frequência esperada de pelo menos cinco casos em cada casa. Outro cuidado a se tomar na aplicação do Q_{xy} é com a heterogeneidade da distribuição dos casos na tabela quádrupla. Uma distribuição muito heterogênea não é indicada para o coeficiente, pois ela já apontaria uma concentração de casos em determinada casa, linha ou coluna. A sugestão é que o cálculo seja realizado sempre que a distribuição dos casos na tabela ficar abaixo de uma relação 70:30 em pelo menos uma das variáveis. Ou seja, não mais de 70% dos casos em uma categoria e não menos de 30% em outra.

No exemplo acima, temos que as mulheres representam 27,4% e os homens 72,6%, no entanto a variável idade divide-se entre 51% abaixo da mediana e 49% acima da mediana, o que cumpre a recomendação de pelo menos uma das variáveis abaixo de 70:30. No exemplo, a distribuição mais homogênea da variável escolaridade “idade” compensou a concentração de casos na categoria sexo, validando o coeficiente de asso-

ciação encontrado. Portanto, podemos ler que ao considerarmos todos os candidatos nas eleições de 2014, as diferenças de idade associadas ao sexo foram praticamente nulas.

2.1.2 Cálculos Adicionais: proporções de Pares Consistentes e Pares Inconsistentes

A interpretação do resultado do coeficiente de associação parte do princípio de que o significado interno do Q_{xy} está ligado à probabilidade de um par de casos diferir em ambos os itens, ou seja, em um ser do sexo masculino e estar acima da mediana da idade ou ser mulher e estar abaixo da mediana de idade. Um par [B, C] na tabela quádrupla (ver quadro 2.1) é chamado de consistente, pois em uma casa ele indica possuir a característica medida nas duas variáveis (X e Y) e na outra apresenta a ausência da característica nas duas variáveis (não-X e não-Y). Já um par [A, D] é chamado de inconsistente, pois em uma variável apresenta a característica analisada e em outra não (X e não-Y) e vice-versa. O primeiro cálculo adicional em uma análise de tabela quádrupla serve para identificar a probabilidade de encontrar pares consistentes nas distribuições. A fórmula para encontrar a probabilidade de pares consistentes é a seguinte:

$$P_c = \frac{2\ x\ (B\ x\ C)}{N^2}$$

Da mesma maneira, a fórmula para encontrar a probabilidade de pares inconsistentes é a seguinte:

$$P_i = \frac{2\ x\ (A\ x\ D)}{N^2}$$

Aplicando as fórmulas ao exemplo acima (tab. 2.1) temos que:

$$P_c = \frac{2\ x\ (B\ x\ C)}{N^2} = \frac{2\ x\ (6.924\ x\ 2.894)}{19.163^2} = \frac{40.076.112}{367.220.569} = \mathbf{0,109}$$

$$P_i = \frac{2\ x\ (A\ x\ D)}{N^2} = \frac{2\ x\ (6.991\ x\ 2.354)}{19.163^2} = \frac{16.456.814}{367.220.569} = \mathbf{0,089}$$

Então, temos que a probabilidade de encontrar pares consistentes é de 0,109 e de inconsistentes de 0,089, portanto, temos proporcionalmente a participação de um pouco mais de pares consistentes em relação aos pares inconsistentes no caso da associação entre sexo do candidato e idade. Essa pequena diferença já é esperada, pois sabemos que a associação entre as duas variáveis é muito baixa.

2.1.3 Cálculos Adicionais: Validade para Inferências

Para o caso das estatísticas inferenciais, outro elemento importante a se considerar quando estamos analisando a força preditiva de um Q_{xy} para associação entre duas variáveis é o tamanho da amostra, ou a forma como as frequências se distribuem nas casas da tabela quádrupla. A recomendação é que existam pelo menos cinco casos esperados se houvesse independência entre as variáveis em cada casa de uma tabela quádrupla. Para saber se a distribuição mínima das frequências esperadas é respeitada sem precisar encontrar o valor para todas as casas, basta multiplicar duas marginais e dividir por N que você terá o Menor Valor Esperado (MVE) para aquela tabela. Para não correr nenhum risco, opte por usar a marginal mais baixa da linha e da coluna. O resultado será o menor valor esperado para uma célula da tabela quádrupla. Portanto, se ficar acima de cinco, todas as demais apresentarão valor esperado superior ao limite mínimo. O cálculo é o seguinte:

$$MVE = \frac{MenorMarginalLinha\ x\ MenorMarginalColuna}{N}$$

Aplicando a fórmula ao nosso exemplo, teríamos que:

$$MVE = \frac{Marg.Não\ X\ x\ Marg.Y}{N} = \frac{5.248\ x\ 9.278}{19.163} = 2.540{,}88$$

Com o resultado de 2.540,88 para “Menor Valor Esperado” não há problemas

neste exemplo, pois estamos muito acima do limite mínimo de casos em cada casa para a realização do teste Q_{xy}. Vamos em frente. Agora, atenção, se a sua tabela quádrupla apresentar um N < 20 o mais provável é que o MVE fique abaixo de cinco e, nesse caso, não devemos calcular o Q_{xy}. Na verdade, essa distribuição só não terá MVE abaixo de cinco se houver uma distribuição totalmente homogênea em cada uma das casas, com frequência igual a cinco em cada uma das quatro casas (5x4 = 20). Porém, aqui também não faz sentido aplicar o Q_{xy}, pois já sabemos que as variáveis são independentes e o Q_{xy} ficará muito próximo a zero.

Muitas vezes, cientistas políticos dispõem apenas de dados amostrais para fazer os testes estatísticos, mas seu objetivo é apresentar resultados que sejam válidos para toda a população, ou seja, fazer inferências. No caso da aplicação inferencial usando o Q_{xy}, é preciso levar em conta um intervalo de confiança para os valores antes de afirmar se a associação encontrada na tabela amostral pode ser extrapolada para toda a população ou não. É evidente que para isso estamos considerando que se trata de uma amostra probabilística. A forma de calcular o intervalo de confiança para inferências a partir do Q_{xy} é outro cálculo adicional da análise de tabelas quádruplas e será apresentada no próximo tópico.

2.1.4 Intervalo de Confiança para o Teste de Correlação Q_{xy}

Até aqui, utilizamos o teste Q_{xy} para indicar a correlação entre duas variáveis, considerando que o número de casos na tabela indica a totalidade, ou seja, o universo estudado. Porém, o coeficiente também pode ser usado em amostras, o que permite passar da estatística descritiva à inferencial, extrapolando os resultados de uma amostra para o universo de casos. Para que isso aconteça, não devemos esquecer que antes de qualquer coisa é preciso que a amostra seja probabilística. Como a teoria da amostragem não é nosso objeto deste capítulo (ver volume I do manual), parto do princípio de que você sabe o que isso significa.

Para podermos inferir resultados de uma amostra ao universo, usamos o conceito de intervalo de confiança (IC), pois ele permite dizer que, em um determinado

intervalo de valores amostrais, há uma probabilidade considerada alta o suficiente de encontrarmos o valor da população na maioria das vezes que extraíssemos amostras dessa população pelos mesmos critérios. Portanto, nosso objetivo é encontrar um valor que seja o limite superior e outro para limite inferior do intervalo de confiança. Ou seja, mínimo e máximo que indicam o intervalo dentro do qual é muito provável encontrar o parâmetro da população. O Intervalo de Confiança mais usado é de 95%, que equivale a dizer que se tirássemos 100 amostras probabilísticas de determinada população, em 95 delas as estatísticas amostrais passariam por um mesmo intervalo.

Não devemos esquecer que a afirmação anterior implica em considerar que em cinco das 100 amostras o valor da estatística estará fora do intervalo de confiança. O valor padronizado do intervalo de confiança de 95% é de z = 1,96. Esse é o número padrão para estabelecer o intervalo de confiança de 95% para resultados amostrais. Com base nele, podemos usar a fórmula abaixo para calcular o limite superior e o inferior do intervalo de valores dentro do qual se deve encontrar o valor da correlação para toda população a partir dos dados obtidos na amostra. Se o resultado indicar que o limite passa pelo valor zero (superior positivo e inferior negativo), então, não podemos dizer que há uma diferença estatística forte o suficiente que permita extrapolar o coeficiente da amostra para a população. Se o intervalo não passar por zero, podemos fazer as inferências estatísticas.

A fórmula para encontrar os limites superior e inferior do intervalo é a seguinte:

Limite superior = Q_{xy} +

$$1{,}96\ x\ \sqrt{\frac{\left(1-\left(Q_{xy}{}^{2}\right)\right)^{2}\ x\ \frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}{4}}$$

Limite inferior = Q_{xy} –

Se estivéssemos trabalhando com uma amostra probabilística no exemplo da relação entre sexo e idade dos candidatos, aplicaríamos a fórmula como segue a partir dos resultados já obtidos:

Limite Superior = 0,098 +

$$1{,}96\ x\ \sqrt{\frac{(1-\ 0{,}098^2)^2\ x\ \frac{1}{6.991}+\frac{1}{6.924}+\frac{1}{2.894}+\frac{1}{2.354}}{4}}\ =\ 1{,}96\ x\ 0{,}000259 = 0{,}00050$$

Limite Inferior = 0,098 –

Assim, os limites seriam:

Limite Superior = 0,098 + 0,00050 = **+ 0,0985**.

Limite Inferior = 0,098 – 0,00050 = **+ 0,0975**.

Os resultados mostram que, se os dados fizessem parte de uma amostra probabilística, o valor do coeficiente de correlação Q_{xy} para a população estaria entre +0,0985 e +0,0975. Como o intervalo não inclui o zero, isso indica que as chances do coeficiente de associação para a população ser nulo não estão dentro do intervalo de confiança que assumimos no início. Portanto, podemos considerar os valores da amostra para inferências à população. De outra forma, se o intervalo passasse por zero, ou seja, o valor do limite superior fosse positivo e do inferior negativo, não poderíamos seguir com as inferências, pois os resultados estatísticos não seriam fortes o suficiente para permitir a extrapolação do coeficiente. O resultado já era esperado, pois ele equivaleria a uma amostra com mais de 19 mil casos, o que é bastante alta para garantir a capacidade de inferência.

Até aqui, utilizamos o coeficiente Q_{xy} para identificar possíveis relações entre duas variáveis dicotômicas dispostas em uma tabela quádrupla. Para tanto, aplicamos os seguintes conceitos:

i) o que estamos testando é a existência de independência entre as categorias de duas variáveis dicotômicas. Se as variáveis forem correlacionadas, identificamos o nível de associação e a direção da mesma – se positiva ou negativa; e,

ii) que podemos testar a relação entre as variáveis em uma população a partir de uma amostra probabilística, desde que os valores dos limites de intervalo de confiança não passem por zero – nesse caso poderemos fazer inferências estatísticas a toda

população.

Além disso, há outra aplicação para o Q_{xy} que é quando o pesquisador insere uma terceira variável dicotômica para testar o quanto essa variável intervém na relação anterior, quando eram consideradas apenas as distribuições de X e Y. A terceira variável, chamada de interveniente ou teste (T), pode interferir de quatro maneiras diferentes na relação anterior, tornando mais rica e detalhada a explicação sobre os fenômenos políticos estudados. Aprenderemos a aplicar o teste Q_{xy} para três variáveis (representado por $Q_{xy:t}$) na próxima seção.

2.1.5 Coeficiente Q_{XY} para três variáveis ($Q_{XY:T}$)

Uma vez realizado o teste de associação para duas variáveis dicotômicas, podemos identificar a existência ou não de relação entre as variáveis, assim como a direção dessa relação, caso exista. Não raras vezes, cientistas políticos pensam em incorporar uma terceira variável à relação para saber se ela é capaz de interferir de alguma forma na associação encontrada anteriormente. Quando a associação é apenas entre duas variáveis, ela é chamada de ordem zero, pois não há nenhum tipo de controle sobre a relação entre as categorias delas. Se controlarmos a relação original por uma terceira variável, essa associação é chamada de ordem um, pois há uma variável controlando a relação identificada entre as duas originais. Portanto, ao acrescentarmos uma terceira variável na medida de associação, temos como resultado uma associação de ordem um ou associação parcial, pois está levando em conta o efeito de controle de uma variável externa.

No tópico anterior e, por tradição, chamamos as variáveis diretamente inseridas no teste de associação de variável X e variável Y. Aqui, a terceira variável será chamada de T (variável teste). Então, a variável teste deve exercer algum efeito sobre a relação entre X e Y, ou seja, ela deve “controlar” de alguma forma a associação medida entre X e Y. Por isso ela também é chamada de variável de controle nos testes de associação.

No tópico anterior, também, as duas variáveis que utilizamos para exemplificar o teste Q_{xy} foram sexo e idade. Testamos se é possível afirmar que candidatos tendem a

ser mais velhos que candidatas. Digamos que agora seja inserida uma terceira variável no teste, ou variável de controle, para verificar como a relação anterior se comporta, dado o efeito da variável teste. No caso, a terceira variável será sucesso na disputa, se ele foi eleito ou derrotado em 2014. Vale lembrar que a variável teste também deve ser dicotômica para o teste $Q_{xy:t}$. Nossa pergunta aqui passa a ser a seguinte: dada a fraca associação já identificada entre sexo e idade, podemos dizer que homens e mulheres apresentam distribuições distintas por idade se considerarmos os que foram eleitos ou não? Ou, em outras palavras, será que há diferenças significativas entre as idades de homens e mulheres eleitos em relação aos homens e mulheres derrotados? Essas variáveis serão utilizadas a seguir para a realização do teste $Q_{xy:t}$ para três variáveis.

Na prática, são realizados dois testes Q_{xy} entre as variáveis X e Y, um para os casos em que há presença da característica da variável teste e outro para os casos em que não há a característica da variável T. Antes do exemplo é preciso discutir quais são os efeitos possíveis da variável teste sobre a relação entre duas variáveis. Basicamente existem quatro possíveis efeitos: a) explicação, b) supressão, c) especificação ou d) sem efeito algum (Davis, 1976). Esses efeitos são constatados quando há alguma diferença entre o coeficiente de associação obtido antes da inserção da variável de controle e depois dela. Havendo alguma diferença, isso é sinal de que a variável de teste exerceu algum tipo de controle sobre os resultados anteriores. A diferença entre os tipos de efeito é:

A) Efeito de Explicação: o efeito de explicação da variável de controle acontece quando o coeficiente de associação de ordem zero é significativo, mas após a inserção da variável teste ele se aproxima de zero. Nesse caso, dizemos que “T explica Y”, pois antes de considerarmos T havia uma relação aparente entre X e Y. Agora, com a inserção de T, a relação inicial é anulada, indicando que a anterior só existia enquanto se desconsiderava a característica de controle (Davis, 1976). É o caso, por exemplo, de termos uma relação significativa entre sexo e idade. Porém, quando inserida a variável teste resultado eleitoral, a relação anterior passa a ser próxima de nula.

B) Efeito de Supressão: Acontece quando o coeficiente de associação parcial, após inserção da variável teste, é mais forte que a associação de ordem zero. Nesse

caso dizemos que T é variável supressiva, visto que ela estava suprimindo a verdadeira relação entre X e Y, que se torna aparente apenas quando há o controle por T (Davis, 1976). Seria o caso de, após inserida a variável resultado eleitoral, a associação sexo e idade apresentar coeficiente maior que o inicial. Diz-se que o efeito é supressivo porque o resultado eleitoral estava escondendo ou suprimindo a verdadeira relação entre X e Y.

C) Efeito de Especificação: esse efeito é diferente dos dois anteriores, quando o coeficiente crescia ou diminuía após a inserção da variável teste. Se considerarmos que uma associação parcial é a combinação de duas associações anteriores, devemos ter em mente a possibilidade de efeitos distintos sobre cada uma das associações anteriores. As diferenças podem ser em termos de magnitude dos coeficientes, assim como até mesmo na inversão dos sinais – com associação positiva em uma e negativa em outra. Quando isso acontece, dizemos que há um efeito de especificação, ou "T especifica XY" (Davis, 1976). Por exemplo, a associação de ordem zero entre sexo e idade poderia ser alta e positiva. Porém, quando inserimos a variável teste resultado eleitoral, podemos encontrar que para homens a relação com idade é significativa e positiva, enquanto que para mulheres ela é não significativa e negativa. Nesse caso, resultado eleitoral está especificando a relação entre sexo e idade.

D) Sem efeito: o quarto tipo de efeito possível é justamente a ausência de qualquer efeito de T sobre a relação XY. Ele é percebido quando o coeficiente de associação entre as duas variáveis na ordem zero é exatamente o mesmo que o obtido após a inserção da variável teste, ou seja, não houve efeito algum da terceira sobre a relação identificada entre as duas anteriores. A ausência de efeito da variável teste é importante para demonstrar que as variáveis X e Y estão realmente associadas entre si, pois a inserção da variável teste não alterou a relação inicial (Davis, 1976). No caso dos exemplos apresentados aqui, equivale a dizer que o coeficiente de correlação entre sexo e idade não se alteraria após a inserção da variável resultado eleitoral.

Como estamos falando, em termos práticos, da repetição do teste Q_{xy} para a presença da característica da variável T e para a ausência da característica na variável T, o que temos é a junção de duas tabelas quádruplas no teste com três variáveis. Ou seja, se no teste entre X e Y tínhamos uma tabela quádrupla (2x2), agora, no teste T, X e Y temos uma tabela óctupla (2x2x2), como o que está representado no quadro a seguir:

Quadro 2.4. Formato das distribuições para $Q_{xy:t}$

		Não Y	Y	TOTAL
T	X	AT	BT	TX
	Não X	CT	DT	$T\overline{X}$
	TOTAL	$T\overline{Y}$	TY	
Não T	X	$A\overline{T}$	$B\overline{T}$	$\overline{T}X$
	Não X	$C\overline{T}$	$D\overline{T}$	$\overline{T}\ \overline{X}$
	TOTAL	$\overline{T}\ \overline{Y}$	$\overline{T}Y$	

Fonte: Davis, 1976

O quadro acima é montado com as casas A, B, C e D repetindo-se duas vezes cada uma, o que comprova que o teste de $Q_{xy:t}$ com três variáveis equivale a dois testes de Q_{xy} simultâneos. Basta calcular um $Q_{xy:t}$ para a tabela quádrupla da parte de cima e outro para a quádrupla da parte de baixo. A letra com o traço acima indica a ausência da característica ($\overline{\ }$ representa "não"). Na parte superior, onde aparece a característica da variável teste, a frequência é representada por T. Na outra, onde não há característica da variável teste, o símbolo é ($\overline{T}$). O mesmo vale para as marginais de X e não-X ($\overline{X}$) e Y e não-Y ($\overline{Y}$). As marginais das linhas apresentam três tipos de pares de valores: com presença de T e X (TX), com presença de apenas uma delas (T $\overline{X}$) ou ($\overline{T}$X) e sem nenhuma das características ($\overline{T}\ \overline{X}$). O mesmo vale para as marginais da variável Y.

Feitas as descrições da tabela óctupla, podemos dizer que existem dois tipos de pares de XY naquela relação: pares ligados a T e pares diferentes de T. Isso porque poderíamos calcular tranquilamente dois coeficientes Q_{xy}, um para pares ligados a T e outro para não ligados a T (quando a característica da variável T não está presente). O princípio por trás do teste $Q_{xy:t}$ para três variáveis é o de que podemos construir coeficientes Q_{xy} para pares ligados a T e para pares sem o efeito de T a partir das localizações no quadro acima. O procedimento é bastante simples e divide-se em três partes:

i) calcula-se o Q_{xy} para pares ligados e o coeficiente Q_{xy} para pares diferentes de T;

ii) calcula-se o Peso para pares ligados e o Peso para pares diferentes;

iii) soma-se o produto do coeficiente Q_{xy} para pares ligados multiplicado pelo Peso para pares ligados com o produto do coeficiente Q_{xy} para pares diferentes multiplicado pelo Peso para pares diferentes.

A fórmula final de $Q_{xy:t}$ é:

$$Q_{xy:t} = \left(Q_{xy}\ ligado\ x\ P.ligados\right) + \ (Q_{xy}\ diferente\ x\ P.diferentes)$$

Onde:

Q_{xy} ligado = coeficiente Q_{xy} para pares ligados a T;

P. ligados = peso para pares ligados a T;

Q_{xy} diferente = coeficiente Q_{xy} para pares diferentes de T;

P. diferentes = peso para pares diferentes.

Assim, o primeiro passo para calcular o $Q_{xy:t}$ para três variáveis é encontrar os valores de Q_{xy} para pares ligados e para pares diferentes de T. As fórmulas, também intuitivas, são as seguintes:

$$Q_{xy}\,ligado = \frac{\left[(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] - \ [\ (AT\ x\ DT) + \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)]}{\left[(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] + \ [\ (AT\ x\ DT) - \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)]}$$

Esta fórmula nos indica qual é o Q parcial para X e Y, controlado por T, ou qual é o Q_{xy} de X e Y em pares ligados a T. Trata-se da melhor forma de predizer a relação entre X e Y quando consideramos apenas os pares ligados a T. Agora, é preciso fazer a mesma coisa para os pares diferentes de T na tabela óctupla. A fórmula é:

$$Q_{xy}\,diferente = \frac{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] - \ [\ \left(AT\ x\ D\overline{\mathrm{T}}\right) + \left(A\overline{\mathrm{T}}\ x\ DT\right)]}{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] + \ [\ \left(AT\ x\ D\overline{\mathrm{T}}\right) + \left(A\overline{\mathrm{T}}\ x\ DT\right)]}$$

O resultado desse cálculo pode ser interpretado como o coeficiente Qxy entre X e Y quando T difere ou o Q entre X e Y para pares diferentes de T. Trata-se da melhor forma de predizer a relação X e Y apenas para os casos em que os pares são diferentes em T. Agora que já encontramos os dois coeficientes, para pares ligados a T e para diferentes de T, o próximo passo é definir os pesos de cada um dos tipos de pares na fórmula final. Com os pesos, nós substituímos uma média simples entre os dois coeficientes

por uma média ponderada pelas diferenças proporcionais dos pares ligados e diferentes de T. Assim, tornamos o coeficiente final mais preciso. As fórmulas são as que seguem:

Peso 1 - ligados: proporção de pares ligados em T entre pares diferentes em X e Y:

$$P1 = \frac{(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right) + (AT\ x\ DT) + \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)}{\left[\left(BT + B\overline{\mathrm{T}}\right) x \left(CT + C\overline{\mathrm{T}}\right)\right] + \left[\left(AT + A\overline{\mathrm{T}}\right) x \left(DT + D\overline{\mathrm{T}}\right)\right]}$$

Peso 2 - diferentes: proporção de pares diferentes em T entre pares diferentes em X e Y:

$$P2 = \frac{\left(BT\ x\ C\overline{\mathrm{T}}\right) + \left(B\overline{\mathrm{T}}\ x\ CT\right) + \left(AT\ x\ D\overline{\mathrm{T}}\right) + \left(A\overline{\mathrm{T}}\ x\ D\mathrm{T}\right)}{\left[\left(BT + B\overline{\mathrm{T}}\right) x \left(CT + C\overline{\mathrm{T}}\right)\right] + \left[\left(AT + A\overline{\mathrm{T}}\right) x \left(DT + D\overline{\mathrm{T}}\right)\right]}$$

À primeira vista, a fórmula assusta, mas no fundo o princípio é simples. Trata-se das frequências de pares T ligados e diferentes nos numeradores do Peso 1 e do Peso 2, respectivamente. O denominador é o mesmo nas duas fórmulas e trata-se do número total de pares diferentes em X e Y. Vamos ver como calcular $Q_{xy:t}$ para três variáveis inserindo como variável teste o resultado eleitoral para todos os candidatos nas eleições de 2014 no cruzamento entre sexo e idade do candidato.

A tabela de cruzamento a seguir apresenta as frequências para cada uma das casas. Para facilitar a identificação dos valores nas casas, são reproduzidos os códigos do quadro 2.4 nas células da tabela.

Tabela 2.2. Cruzamento entre sexo e idade de candidato controlado por resultado eleitoral

RESULTADO	SEXO	IDADE		TOTAL
		Abaixo Mediana (Não Y)	Acima Mediana (Y)	
Eleito (T)	Homem (X)	557 (AT)	765 (BT)	1.322 (TX)
	Mulher (Não X)	66 (CT)	97 (DT)	163 ($T\overline{X}$)
	Total	623 ($T\overline{Y}$)	862 (TY)	1.485
Não Eleito (não T)	Homem (X)	6.434 ($A\overline{T}$)	6.159 ($B\overline{T}$)	12.593 ($\overline{T}X$)
	Mulher (Não X)	2.827 ($C\overline{T}$)	2.257 ($D\overline{T}$)	5.084 ($\overline{T}\ \overline{X}$)
	Total	9.261 ($\overline{T}\ \overline{Y}$)	8.416 ($\overline{T}Y$)	17.677

Fonte: autor com dados do TSE.

Aplicando as fórmulas, começamos por encontrar os coeficientes Q para pares ligados a T:

$$Q_{xy}\,ligado = \frac{\left[(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] - \left[\,(AT\ x\ DT) + \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)\right]}{\left[(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right)\right] + \left[\,(AT\ x\ DT) - \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)\right]}$$
$$= \frac{[(765\ x\ 66) + (6159x\ 2827)] - [(557\ x\ 97) + (6434\ x\ 2257)]}{[(765\ x\ 66) + (6159x\ 2827)] + [(557\ x\ 97) - (6434\ x\ 2257)]} = \frac{2886416}{2994474}$$
$$= \mathbf{0,963}$$

Agora, a fórmula para encontrar o coeficiente Q para pares diferentes de T:

$$Q_{xy}\,diferente = \frac{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] - \left[\,\left(AT\ x\ D\overline{\mathrm{T}}\right) + \left(A\overline{\mathrm{T}}\ x\ DT\right)\right]}{\left[\left(BT\ x\ C\overline{\mathrm{T}}\right) + \left(B\overline{\mathrm{T}}\ x\ CT\right)\right] + \left[\,\left(AT\ x\ D\overline{\mathrm{T}}\right) + \left(A\overline{\mathrm{T}}\ x\ DT\right)\right]} =$$
$$= \frac{[(765\ x\ 2827) + (6159\ x\ 66)] - [(557\ x\ 2257) + (6434\ x\ 97)]}{[(765\ x\ 2827) + (6159\ x\ 66)] + [(557\ x\ 2257) + (6434\ x\ 97)]} = \frac{687902}{4450396}$$
$$= \mathbf{0,154}$$

O próximo passo é encontrar os pesos para cada grupo de pares. Começamos pelo peso dos pares ligados a T, chamado P1:

$$P1 = \frac{(BT\ x\ CT) + \left(B\overline{\mathrm{T}}\ x\ C\overline{\mathrm{T}}\right) + \left(AT\ x\ DT\right) + \left(A\overline{\mathrm{T}}\ x\ D\overline{\mathrm{T}}\right)}{\left[\left(BT + B\overline{\mathrm{T}}\right) x \left(CT + C\overline{\mathrm{T}}\right)\right] + \left[\left(AT + A\overline{\mathrm{T}}\right) x \left(DT + D\overline{\mathrm{T}}\right)\right]}$$
$$= \frac{(765\ x\ 66) + (6159\ x\ 2827) + (557\ x\ 97) + (6434\ x\ 2257)}{[(765 + 6159)\ x\ (66 + 2827)] + [(557 + 6434)\ x\ (97 + 2257)]} = \frac{32037550}{36487946}$$
$$= \mathbf{0,878}$$

Repete-se o procedimento para encontrar o peso dos pares diferentes de T, o P2:

$$P2 = \frac{\left(BT\ x\ C\overline{\mathrm{T}}\right) + \left(B\overline{\mathrm{T}}\ x\ CT\right) + \left(AT\ x\ D\overline{\mathrm{T}}\right) + \left(A\overline{\mathrm{T}}\ x\ DT\right)}{\left[\left(BT + B\overline{\mathrm{T}}\right) x \left(CT + C\overline{\mathrm{T}}\right)\right] + \left[\left(AT + A\overline{\mathrm{T}}\right) x \left(DT + D\overline{\mathrm{T}}\right)\right]}$$
$$= \frac{(765\ x\ 2827) + (6159\ x\ 66) + (557\ x\ 2257) + (6434\ x\ 97)}{[(765 + 6159)\ x\ (66 + 2827)] + [(557 + 6434)\ x\ (97 + 2257)]} = \frac{4450396}{36487946}$$
$$= \mathbf{0,121}$$

Agora já temos todos os fatores necessários para o cálculo do coeficiente Q para as três variáveis. A fórmula a ser aplicada é:

$$Q_{xy:t} = \left(Q_{xy}\ ligado\ x\ P.ligado\right) + \ (Q_{xy}\ diferente\ x\ P.diferente)$$

$$Q_{xy:t}\ =\ (0{,}963\ x\ 0{,}878)\ +\ (0{,}154\ x\ 0{,}121)$$

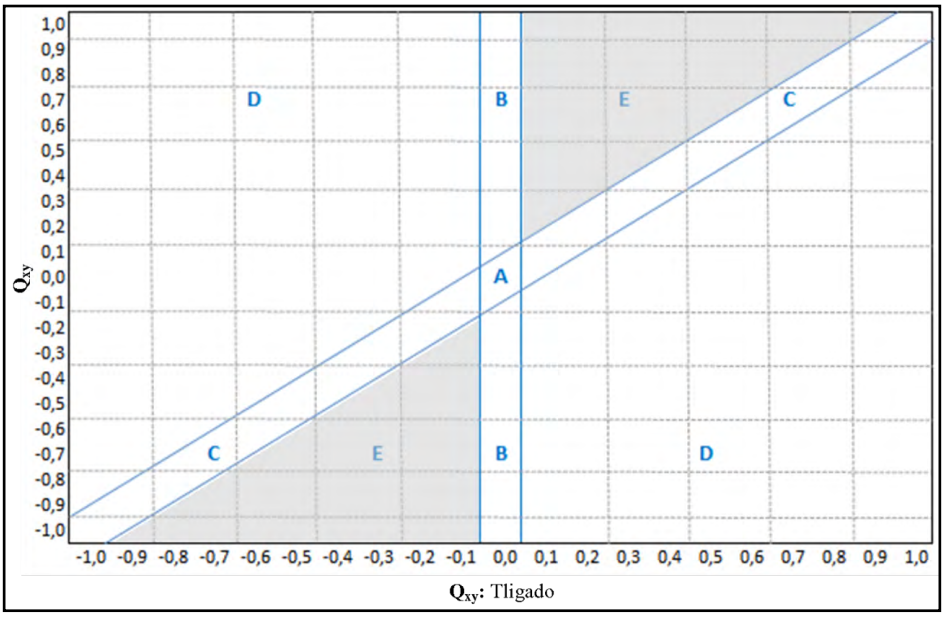
$$Q_{xy:t}\ =\ 0{,}845\ +\ 0{,}018$$

$$Q_{xy:t}\ =\ \mathbf{0,864}$$

Resultado: o coeficiente de associação $Q_{xy:t}$ para sexo e idade do candidato, controlada pelo resultado eleitoral é de 0,864 ou 86,4%. Como o coeficiente é positivo, significa que há uma associação de 86,4% entre ser homem mais velho, quando controlado pelo sucesso eleitoral. Perceba que ao compararmos com o coeficiente do Q_{xy} sem a variável de controle há uma grande diferença. Agora, com controle, o coeficiente ficou mais elevado. Quando consideramos apenas sexo e idade para todos os candidatos, a associação ficou em apenas 9,8%. Mas quando controlada pelo resultado eleitoral, subiu para 86,4%. Essa diferença é o que indica o efeito da variável teste sobre a relação de ordem zero. A questão agora é saber se a magnitude da diferença é suficiente para garantir que a variável teste interferiu de fato na associação de ordem zero.

Podemos entender o $Q_{xy:t}$ para três variáveis como uma média ponderada do Q_{xy} parcial e do Q_{xy} diferencial, sendo os pesos as proporções de pares ligados em T e diferentes de T. Um dos mais importantes princípios do teste de associação $Q_{xy:t}$ com três variáveis é que, qualquer que seja o valor da associação na ordem zero, o valor parcial poderá assumir também qualquer valor entre os limites teóricos +1,00 e -1,00. Esse princípio nos permite estabelecer relações possíveis para a análise de três variáveis em um espaço bidimensional, onde o eixo Y representa o valor que Q_{xy} e o eixo X representa os valores de $Q_{xy:t}$, ambos podendo variar nos limites teóricos de +1,00 a -1,00. Em um artigo publicado em 1950, Kendall e Lazarsfeld estabeleceram os limites das divisões desse gráfico de ordenadas, chegando a cinco regiões significativas, como indicado no gráfico a seguir.

Gráfico 2.1. Áreas de distribuições de posições dos coeficientes de ordem zero e parcial



Fonte: reproduzido de Kendal e Lazarsfeld, 1950.

A região A indica a área onde se situam as associações muito próximas a zero tanto na correlação de ordem zero, quanto na parcial. Ou seja, não há relação entre X e Y com ou sem o controle de T;

As duas regiões B indicam os resultados em que a relação de ordem zero é significativa, porém, a associação parcial fica próxima a zero, quer dizer, a variável de controle explica a relação entre X e Y, pois o coeficiente era alto (positivo ou negativo) antes do controle e caiu após a inserção da variável teste;

As regiões C, que seguem a diagonal dos dois eixos, indicam os casos em que a correlação de ordem zero e a correlação parcial apresentam resultados muito próximos. Os coeficientes que caem nessa região apresentam uma correlação entre X e Y que independe do controle da variável T. Portanto, a variável teste não exerce efeito significativo sobre a associação de ordem zero.

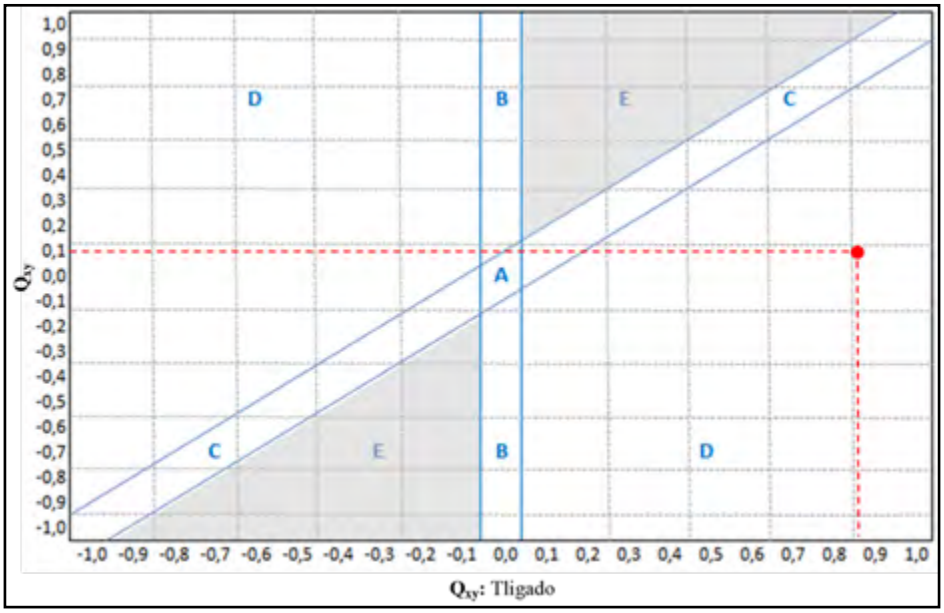
Nas regiões D encontram-se os resultados em que o coeficiente parcial é maior que o coeficiente de ordem zero – com o mesmo sinal ou com sinal oposto. Dá-se quando a relação entre X e Y controlada por T aumenta na mesma direção ou mudando de sinal. Aqui há um efeito evidente da terceira variável sobre a relação de ordem zero.

As regiões E indicam os casos em que tanto a correlação de ordem zero

quanto a parcial são significativas. Porém, o coeficiente da correlação parcial é menor que o coeficiente de ordem zero. Nesses casos há uma oscilação entre poder explicar a relação e não poder explicá-la.

Conhecendo o gráfico de distribuição dos coeficientes de ordem zero e parcial, temos condições de verificar se o nosso coeficiente parcial (com variável de controle resultado eleitoral) exerce algum tipo de influência sobre o coeficiente de ordem zero (associação entre sexo e idade dos candidatos em 2014). O resultado está plotado no gráfico 2.2 a seguir:

Gráfico 2.2. Interseção entre Q_{xy} e $Q_{xy:t}$ para associação de ordem zero e parcial anterior



Fonte: reproduzido de Kendal e Lazarsfeld, 1950.

O gráfico 2.2 mostra que o ponto de interseção entre os dois coeficientes no gráfico de ordenadas X e Y fica na região D, indicando que a diferença entre o coeficiente de ordem zero e o coeficiente parcial é grande e nos permite dizer que o controle alterou significativamente a associação de ordem zero. Isso já era possível antecipar ao comparar os dois coeficientes, pois ele salta de 9,8% para 86,4% quando controlado pelo resultado eleitoral. Em outras palavras, considerando os candidatos às eleições nacionais de 2014, a distribuição de idades não apresenta associação com sexo quando consideramos todos os candidatos. Porém, quando controlamos a relação pelo resul-

tado eleitoral, o coeficiente de associação controlada passa a ser alto, indicando que, entre os eleitos, os homens tendem a ser mais velhos que as mulheres. O que reduzia o coeficiente de associação de ordem zero é que entre os derrotados a associação é muito baixa, tendendo a existir tanto homens abaixo quando acima da mediana de idade. Já entre apenas os eleitos, os homens tendem a ter mais idade que as mulheres.

2.2 Teste com variáveis ordinais - Coeficiente Gama (G)

Quando a tabela tem mais que duas colunas e duas linhas, deixa de ser tabela quádrupla e passa a ser classificada pelo número de linhas e colunas, chamada de "LxC", ou, linhas por colunas. Uma tabela com três linhas e quatro colunas é denominada de "3x4". Lembrando que cada linha representa o número de categorias válidas da variável X e cada coluna representa o número de categorias válidas da variável Y. Até aqui, falamos de coeficientes que medem a força da relação entre duas variáveis dicotômicas, ou seja, em tabelas "2x2". A partir de agora, abordaremos os coeficientes usados em tabelas com mais de duas linhas ou colunas, quer dizer, para aquelas variáveis que apresentam três ou mais categorias. E pelo menos uma das variáveis é ordinal.

O número de categorias não é suficiente para definir o tipo de teste, que depende do nível de mensuração das variáveis. O Coeficiente Gama (G) é um teste indicado para medir o grau de relação entre duas variáveis categóricas ordinais ou pelo menos uma ordinal e outra nominal. As variáveis ordinais apresentam como característica um arranjo de categorias que é "transitivo", ou seja, ela nos dá uma noção de ordem entre as categorias (Davis, 1976). Há uma transição entre a categoria de menor valor, para a de maior valor, diferente de uma variável nominal, que apenas separa grupos por semelhanças ou diferenças de características. Em uma variável ordinal, as categorias podem representar valores de mais negativos para mais positivos. Por exemplo, uma variável categórica é o nível de escolaridade das pessoas. Sabemos que uma pessoa que tem nível de escolaridade Fundamental passou menos tempo na escola do que alguém que tem nível Médio, que por sua vez tem menos escolaridade do que aqueles com nível Superior. Nesse caso, o nível Médio faz a transição entre o fundamental e o superior.

Outro exemplo é tamanho dos municípios por categorias: micro, pequeno, médio e grande. Há uma transitividade no número crescente de habitantes da primeira para a última categoria dessa variável.

Já vimos que qualquer variável pode ser dicotomizada para se verificar o nível de associação pelo teste Q_{xy}. Então, por que optar por um teste de associação entre variáveis com mais de duas categorias? A resposta é: quando agregamos categorias para dicotomizar as variáveis, corremos o risco de esconder pares consistentes em conjuntos de categorias com predomínio de pares inconsistentes. Isso diminui a precisão dos resultados de testes como o Q_{xy}. Se usarmos testes com variáveis cujas categorias não foram agregadas, a precisão pode aumentar.

O cálculo para o coeficiente Gama segue o mesmo princípio do Q_{xy}. Ele compara o volume de pares consistentes com o de pares inconsistentes, dividindo pelo total de pares diferentes em X e Y. Só que o Gama faz isso para todas as categorias que aparecem nas variáveis, aumentando o detalhamento das relações entre elas. Como o coeficiente Gama aplica o mesmo princípio para todos os pares de valores, a fórmula do seu cálculo não é tão simples (produtos cruzados) como a do Q_{xy}. Um detalhe importante para o teste Gama quando as duas variáveis são ordinais é que as categorias devem ser organizadas na tabela de forma que a linha superior e a coluna da direita correspondam aos extremos positivos das categorias das duas variáveis, conforme o esquema a seguir. Nele, a categoria que representa o valor mais alto de X(++) está na primeira linha da tabela e a categoria que representa o valor mais alto de Y(++) está na coluna da direita.

Quadro 2.5. Organização das categorias ordinais na tabela de contingência para teste G

Var X	Var Y			Total
	Y--	Y-+	Y++	
X++			→	
X -+				
X --				
Total				

Vejamos como calcular o coeficiente Gama para associação entre as variáveis i) tamanho do município em número de eleitores em 2016 segundo TSE e ii) escolarida-

de do prefeito eleito em três categorias (lê e escreve e fundamental, médio e superior), de acordo com o declarado ao TSE. Para evitar distorções nas comparações, consideraremos apenas os municípios com eleição em turno único, ou seja, com até 200 mil eleitores. Assim, distribuímos os municípios aproximadamente por quartis, em ordem crescente de eleitores, formando quatro categorias. As perguntas que movem esse teste poderiam ser: será que existe alguma relação entre escolaridade do eleito e tamanho do município? Em municípios maiores devemos encontrar uma concentração de prefeitos com maior escolaridade? O objetivo é identificar se existe ou não associação entre escolaridade do eleito e tamanho do município. A hipótese nula é que não existe. A hipótese alternativa é que as duas variáveis estão associadas. Como elas são ordinais, deve-se usar o Coeficiente Gama. Se o resultado por positivo, teremos uma associação positiva entre nível de escolaridade e tamanho do município; se for negativo, a associação será inversa – prefeitos com menor escolaridade tendem a se concentrar nos maiores municípios. Se o coeficiente for zero ou próximo dele, a associação será muito baixa, ou seja, não há relação entre escolaridade do prefeito e tamanho do município.

Tabela 2.3. Cruzamento entre escolaridade do prefeito eleito em 2016 e Tamanho do município

ESCOLARIDADE DO PREFEITO ELEITO	Tamanho em número de eleitores				TOTAL
	Até 5 mil	De 5 a 10 mil	De 10 a 20 mil	De 20 a 200 mil	
Superior	706	779	799	866	3.150
Médio	516	496	350	226	1.588
Lê, escreve e fundamental	319	215	139	77	750
TOTAL	1.541	1.490	1.288	1.169	5.488

Fonte: autor a partir do TSE

Seguindo a exigência apresentada no quadro 2.5, a primeira linha da tabela contém os dados da escolaridade mais alta (escolaridade superior) e segue até a mais baixa (lê, escreve ou nível fundamental). E as colunas começam da categoria mais baixa à esquerda para subir em direção à direita, iniciando em municípios com até 5 mil habitantes até chegar ao quartil mais alto, entre 20 e 200 mil habitantes. A fórmula para calcular o coeficiente Gama para a relação entre as categorias é:

$$G = \frac{PC - PI}{PC + PI}$$

Onde:

PC = Produto cruzado total para pares consistentes;

PI = Produto cruzado total para pares inconsistentes.

O cálculo do PC e do PI é um pouco trabalhoso, mas nada complicado. Para encontrar o Produto cruzado total para os pares consistentes (PC), faça o seguinte:

a) Comece multiplicando a frequência da célula superior direita com cada uma das frequências situadas abaixo e à direita dela. No nosso exemplo, seria 866. Esse valor será multiplicado seis vezes, pois há seis casas à esquerda e abaixo dela.

b) Em seguida repita o procedimento para a célula que se encontra imediatamente à esquerda da anterior (no exemplo, o valor é 799) multiplicando-a por todas as frequências situadas abaixo e à esquerda dela. São quatro casas.

c) Repita o mesmo procedimento até chegar à penúltima coluna, pois não havendo nenhum valor à esquerda e abaixo da última coluna à esquerda, ela deve ser desconsiderada. A soma de todas essas multiplicações é o PC.

Para o cálculo dos Pares Inconsistentes (PI), produto cruzado total para os pares inconsistentes, os passos são parecidos, porém, no sentido inverso:

a) Multiplique a frequência do canto superior esquerdo (706) com todas as frequências abaixo e à direita dela.

b) Repita o mesmo procedimento para a primeira frequência da segunda coluna à esquerda (779) e assim sucessivamente.

c) repita o procedimento até a penúltima coluna, pois não havendo frequência à direita e abaixo da última coluna, ela deve ser desconsiderada. A soma de todas essas multiplicações é o PI.

A tabela 2.4 a seguir apresenta todas as sequências de multiplicações para pares consistentes e inconsistentes, além das somas dos valores que serão aplicados à fórmula do coeficiente Gama:

Tabela 2.4. Cálculos para Pares Consistentes e Inconsistentes

ESCOLARIDADE DO PREFEITO ELEITO	Tamanho em número de eleitores				TOTAL
	Até 5 mil	De 5 a 10 mil	De 10 a 20 mil	De 20 a 200 mil	
Superior	706	779	799	866	3.150
Médio	516	496	350	226	1.588
Lê, escreve e fundamental	319	215	139	77	750
TOTAL	1.541	1.490	1.288	1.169	5.488
P. Consistentes	446.856	429.536	303.100		
	276.254	186.190	120.374		
	412.284	396.304			
	254.881	171.785			
	401.964				
	248.501				
	72.094	48.590	31.414		
	75.250	111.650			
	158.224				
				Soma	**4.145.251**
P. Inconsistentes		350.176	247.100	159.556	
		151.790	98.134	54.362	
			272.650	176.054	
			108.281	59.983	
				180.574	
				61.523	
		110.940	71.724	39.732	
			68.944	38.192	
				26.950	
				Soma	**2.276.665**

Antes mesmo de aplicarmos a fórmula, podemos perceber que o montante obtido em pares consistentes (PC) é bem superior ao montante de pares inconsistentes (PI). Isso nos permite antecipar que há alguma associação entre as duas variáveis, pois a soma dos pares consistentes resulta em quase o dobro do valor dos pares inconsistentes. Além disso, podemos dizer que a associação é positiva, pois os consistentes são em maior número que os inconsistentes. Aplicando a fórmula, temos que o coeficiente Gama é:

$$G = \frac{PC - PI}{PC + PI} = \frac{4.145.251 - 2.276.665}{4.145.251 + 2.276.665} = \frac{1.868.586}{6.421.916} = \mathbf{0,290}$$

O resultado é um coeficiente de associação Gama de +0,290 (+29,0%) entre escolaridade do prefeito eleito e tamanho do município. Portanto, conforme aumenta o tamanho do município, maior tende a crescer a escolaridade do prefeito eleito. O teste mostra uma associação moderada entre escolaridade do prefeito e tamanho do município, o que pode ser lido como municípios menores tendem a eleger prefeitos com menor escolaridade.

O nível de detalhamento do coeficiente depende do número de categorias nas variáveis. Se, para reduzirmos o trabalho, agregarmos duas categorias em uma para tornar o teste mais simples, o resultado será uma redução na precisão do coeficiente. Essa redução depende do grau de "ocultação" de pares consistentes que acontecerá quando as categorias forem agregadas. Como tem mais de duas categorias, as variáveis ordinais usadas no cálculo do coeficiente Gama permitem resultados mais refinados do que ao usarmos um coeficiente Q_{xy} para variáveis dicotômicas. O fato é que se a tabela de contingência apresentar uma distribuição dos casos muito distinta entre as categorias vizinhas, a agregação das mesmas em variável dicotômica pode "esconder" importantes diferenças que deixarão de existir no cálculo do coeficiente. Nesses casos, apesar da maior dificuldade para o cálculo, recomenda-se o uso do Gama sem agregar categorias.

2.3 Referências bibliográficas do Capítulo II

Davis, J. A. (1976). *Levantamento de Dados em Sociologia: uma análise estatística elementar*. Rio de Janeiro: Zahar Editores.

Kendall, P. L., & Lazarsfeld, P. F. (1950). Problem of Survey Analysis. In Merton, R., & Lazarsfeld, P. F. (Orgs.). *Continuities in Social Research*. New York: Free Press.

Yule, G. U.; & Kendall, M. G. (1937). *An introduction to the theory of statistics*. London: Charles Griffin.

2.4 Exercícios propostos do Capítulo II

2.4.1 Considerando a importância das coligações eleitorais em sistemas multipartidários, a partir das tabelas de cruzamentos bivariados entre estar ou não coligado e ser eleito prefeito municipal em 2016, faça os seguintes testes:

a) teste de independência de Q_{xy};

b) calcule as proporções de Pares Consistentes e Pares Inconsistentes;

c) calcule o Menor Valor Esperado (MVE) para o cruzamento;

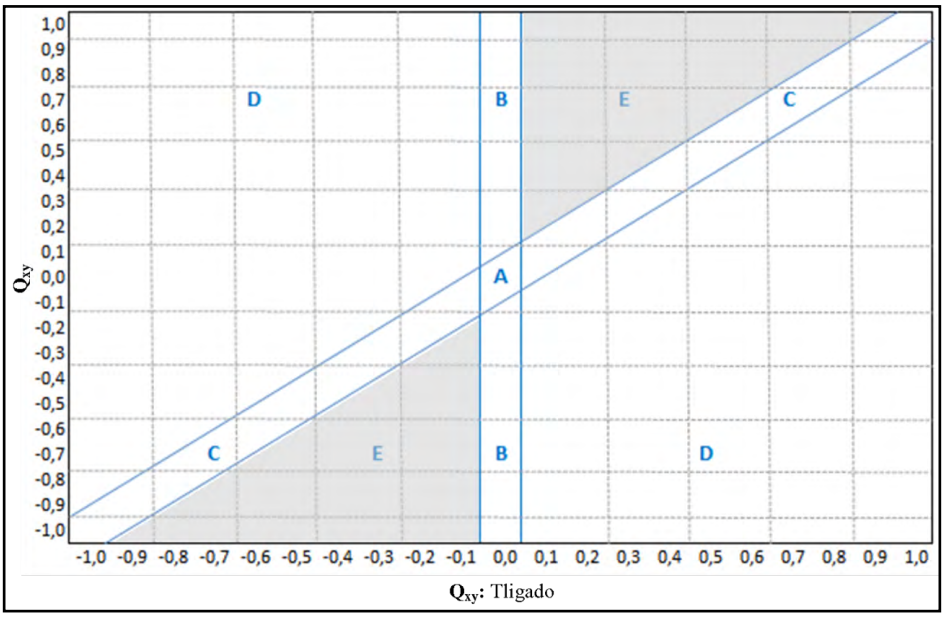
d) considerando que a tabela representasse dados de uma amostra, use os valores para identificar o intervalo de confiança.

TIPO CAND.	NÃO ELEITO (Não Y)	ELEITO (Y)	TOTAL
COLIGADO (X)	8.208	5.410	13.618
SEM COLIGAR (Não X)	2.445	112	2.557
TOTAL	10.653	5.522	16.175

2.4.2 Acrescente a variável Sexo do candidato (Mulher ou Homem) como controle para a associação entre ser eleito e estar ou não coligado. Faça o teste com controle e analise as diferenças para os casos de candidatas por coligações ou por partidos isolados quanto ao resultado eleitoral.

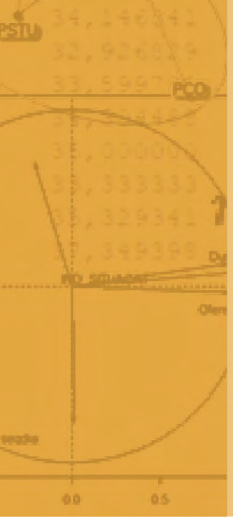
SEXO	TIPO CAND.	NÃO ELEITO (Não Y)	ELEITO (Y)	TOTAL
MULHER (T)	PART. COLIGADO (X)	1.164	594	1.758
	PART. ISOLADO (Não X)	254	8	262
	TOTAL	1.418	602	2.020
HOMEM (Não T)	PART. COLIGADO (X)	7.044	4.816	11.860
	PART. ISOLADO (Não X)	2.191	104	2.295
	TOTAL	9.235	4.920	14.155

Em seguida, as linhas para identificar o ponto de interseção entre os coeficientes de ordem zero e com variável teste no gráfico a seguir. Interprete o resultado em função da área em que se encontrar o ponto de interseção.



2.4.3 Calcule o coeficiente Gama para a associação entre as duas variáveis ordinais "escolaridade do candidato" e "votação em categorias" a partir dos dados da tabela abaixo, que consideram todos os candidatos a prefeito que disputaram as eleições de 2016 no Brasil. Interprete o resultado. Lembre-se de considerar sempre a magnitude e o sinal do coeficiente no momento de interpretá-lo.

ESCOLARIDADE	CATEGORIA DE VOTAÇÃO				TOTAL
	Abaixo de 10%	De 10 a 30%	De 30% a 50%	Acima de 50%	
SUPERIOR	2.233	1.620	3.229	2.407	9.489
MEDIA	1.053	657	1.662	1.274	4.646
FUNDAMENTAL	489	284	779	601	2.153
TOTAL	3.775	2.561	5.670	4.282	16.288



Capítulo III

Análise de Dados Categóricos

O objetivo da modelagem é facilitar a interpretação de fenômenos empíricos, reduzindo a complexidade deles, e não esterilizar a interpretação da realidade.

Nos dois primeiros capítulos, discutimos testes de associação e análises de pares de categorias em tabelas de contingência sem a necessidade de *softwares* estatísticos. Agora, passaremos a uma categoria distinta de análises, nem melhor, nem mais elaborada, que apenas cumpre objetivos diferentes. Para o uso de teste em tabelas de contingência ou bancos de dados com centenas ou milhares de casos, os pacotes estatísticos são úteis porque economizam tempo e energia. Além do mais, quando se está trabalhando com múltiplas variáveis e centenas de casos, as relações entre eles tornam-se tão complexas que passa a ser indispensável alguma forma de redução de dimensões. E esse é o principal objetivo das análises multivariadas que serão apresentadas aqui, a redução de dimensões para um número que permita a análise de relações. Começaremos apresentando alguns conceitos básicos sobre o que são dados categóricos à luz dos métodos quantitativos e quais as diferenças deles para os dados contínuos. Ainda assim, alguns testes apresentados neste capítulo podem ser usados com variáveis contínuas. Para todas as análise no capítulo, usaremos a interface *RCommander* do pacote estatístico de código aberto, R, já apresenta-

do no primeiro volume deste Manual. Além disso, para alguns testes, foi necessária a incorporação de um plug-in ao *RCommander* chamado *FactoMineR*. Para mais detalhes sobre instalação e uso do plug-in consultar Lé *et al.* (2008).

Para começar, é preciso apresentar um teste de confiabilidade de medidas, que garante alguma validade para a transformação léxica de conceitos em variáveis empíricas complexas, como índices formados por variáveis categóricas. Em seguida, passaremos para o primeiro teste propriamente dito, uma análise bivariada de correspondência entre variáveis nominais. A partir de então, chegaremos aos testes multivariados. Iremos da análise múltipla de correspondência para a análise de componentes principais, usada tanto em variáveis categóricas quanto em contínuas. Por fim, serão apresentadas as principais características das análises de agrupamentos, os chamados testes de *cluster*. Enquanto a correspondência busca identificar proximidades e distâncias espaciais entre variáveis e unidades, os componentes principais partem de vetores teóricos não correlacionados com as variáveis originais, e a análise de agrupamento busca identificar que categorias de diferentes variáveis mais se aproximam entre si. Em todos os casos, o princípio é a distribuição espacial euclidiana.

3.1 Análise de Dados Categóricos

A análise de dados categóricos faz com que informações qualitativas a respeito de eventos ou objetos pesquisados sejam tratadas e analisadas a partir de técnicas quantitativas. Aqui, dado qualitativo e variável categórica serão usados como sinônimos. A definição dada a elas é de que um dado qualitativo é uma representação atribuída a quantidades de manifestações de determinada qualidade. Então, chamamos de "variável categórica" a característica medida em determinado objeto de estudo que apresenta duas ou mais variações em quantidades distintas. O dado qualitativo classifica, assim, um fenômeno quase que imponderável a partir de premissas ontológicas e semânticas. Por exemplo: "partido de direita" e "partido de esquerda" são construções semânticas, pois não existem na realidade. São construídos em função de comportamentos distintos que seus agentes políticos adotam em relação aos mesmos temas. Porém, com

essa classificação é possível instrumentalizar o reconhecimento do evento, analisar seu comportamento e suas relações com outros eventos. Nesse sentido, trata-se de uma qualificação normativa que dá um caráter objetivo à análise. Estar em determinado partido pode ser entendido como uma qualidade. Ter sido eleito em uma eleição é outra qualidade que se opõe à qualidade de ter sido derrotado. A análise de dados qualitativos com técnicas quantitativas é considerada uma alternativa à pesquisa qualitativa, que se ocupa dos mesmos eventos, porém com menor restritividade técnica e maior possibilidade de intervenção da subjetividade do pesquisador.

Antes de começar a medir características qualitativas é preciso ter em mente a distinção entre **objeto** e **atributo**. As técnicas qualitativas usam uma estratégia de mensuração de atributos, ou seja, a mensuração não se dá sobre o objeto em si, a coisa, mas sobre uma ou algumas de suas características e predicados, aqui chamados de atributos. O que pretendemos aqui é analisar atributos dos objetos e não os objetos em si. Existem dois tipos de medidas que servem para identificar esses atributos: as **fundamentais** e as **derivadas**. As medidas fundamentais são aquelas em que a mensuração é feita diretamente sobre o objeto. Por exemplo: quando se usa uma balança para medir o peso das pessoas. Estamos medindo o atributo diretamente no objeto. Em outro caso, nas medidas derivadas, é feita uma projeção a partir de uma medida indireta e não diretamente sobre o objeto. Por exemplo: a opinião sobre preconceito medida a partir das respostas em um *survey*. Cada respondente emite sua opinião sobre comportamentos preconceituosos, mas o pesquisador não tem certeza se o respondente adota o comportamento que ele diz preferir ou não. A opinião é uma medida derivada a respeito de comportamento sobre determinado tema. As técnicas de análise de dados categóricos em ciência política preocupam-se com a análise de atributos dos objetos a partir de medidas derivadas.

As representações de relações entre duas variáveis categóricas podem ser feitas por gráficos ou tabelas (para mais detalhes, ver tópico específico sobre representação gráfica de dados no volume I deste Manual). A recomendação geral é que a análise de dados qualitativos dê-se a partir de representações visuais, como gráficos, em lugar de tabelas, pois o que se busca é a redução de dimensionalidades. Depois de observar toda a complexidade das variáveis a partir de medidas já discutidas, o pesquisador

precisa ter uma medida de relação geral que lhe permita tirar alguma conclusão. Nos próximos tópicos, serão apresentados alguns testes e indicadores estatísticos utilizados em pesquisas na área de ciência política para a produção de informações sobre variáveis categóricas. Ou seja, a respeito de atributos dos objetos analisados. Para tanto, seguiremos a seguinte ordem: a produção de indicadores para a análise univariada e o impacto de uma variável sobre determinados fenômenos. Em seguida, são apresentados alguns testes estatísticos multivariados comuns para dados categóricos.

3.2 Teste de Confiabilidade para Indicadores Estatísticos

Uma das formas de medir as variações de qualidades em objetos é pela reunião de características que se complementam ou que agregam informações em uma mesma direção. São os chamados índices, já discutidos no volume I deste manual. A criação de indicadores ou índices é uma das formas mais conhecidas para redução das dimensões de uma ou mais variáveis, ou seja, ela reduz categorias através da contagem de impacto da presença de determinada característica na variável. Essa redução dimensional permite que a nova variável seja mais informativa que seus componentes individuais. No entanto, ao reunir informações em um índice, estamos reduzindo dimensões e perdendo detalhes. O pesquisador precisa levar isso em consideração ao decidir pela agregação de variáveis. Além do mais, é preciso considerar se as variáveis isoladas, quando reunidas, estão mesmo produzindo uma nova variável que seja informativa e que transmita aquilo que se espera dela. Uma coisa é a definição semântica, outra é o comportamento das variáveis na realidade. Para testar se a junção de variáveis em um índice justifica-se, existem os testes de confiabilidade. Esse tipo de teste gera um coeficiente que indica se as variáveis usadas para a criação do índice estão mesmo contribuindo entre si ou se, ao contrário do que se pensava inicialmente, não agregam informação nova em relação ao que já temos com as variáveis isoladas (Cronbach, 1971).

Vejamos um exemplo. Pesquisas sobre o conteúdo do Horário Gratuito de Propaganda Eleitoral (HGPE), desenvolvidas pelo CPOP a partir da metodologia proposta pelo Doxa/Iesp produzem informações sobre variáveis isoladas que uma vez reunidas

podem gerar novos índices. É o caso de um conjunto de variáveis que pretendem medir as estratégias típicas de desafiante, aquelas que são usadas no HGPE de candidatos que, espera-se, tenham comportamento de opositor. Para exemplificar aqui, usaremos o banco de dados do HGPE dos candidatos à presidência da República em 2014. Nele, existem três variáveis binárias que indicam oposicionismo (presença / ausência da característica): "Apelo a mudanças", "Ataques à administração em curso" e "Ataques a adversários". Poderíamos pensar em criar um índice de oposicionismo no HGPE reunindo essas três variáveis em uma única, que contaria com três categorias de oposição: Alta (quando há presença para as três variáveis), Média (quando há presença apenas para duas delas), Baixa (quando só há presença em uma delas). E quando as três variáveis forem zero, teremos ausência de oposicionismo no HGPE. Ainda que o normal seja criar a nova variável, somando valores diretamente, recomenda-se testar a confiabilidade da medida, ou seja, verificar se em conjunto essas três informações estão contribuindo de fato para a explicação do fenômeno ou não.

O teste de confiabilidade mais usado é o *α de Cronbach*. Nele, são consideradas as covariações conjuntas das variáveis para verificar a partir dos desvios se as mudanças das categorias se dão de forma semelhante e na mesma direção. Para medir a consistência ou confiabilidade de determinado indicador é usado o coeficiente *α de Cronbach*, que deve ser interpretado como um coeficiente de correlação ao quadrado (r^2) com média supostamente real a respeito do fenômeno estudado (Pereira, 2004). Apesar disso, não devemos confundir o *α de Cronbach* com um coeficiente de correlação. Ele não indica correlação entre os valores. Se o valor do coeficiente é alto ou baixo depende do objeto pesquisado. Pode-se tentar incluir ou excluir alguma variável das já consideradas para testar possíveis alterações no coeficiente de *α de Cronbach*, caso o valor não seja satisfatório.

É importante observar que o teste parte do pressuposto de que todas as variáveis têm correlações positivas entre si. Portanto, antes de produzi-lo, é preciso verificar uma matriz de correlação com as variáveis inseridas no indicador. Se o coeficiente de correlação for positivo significa que as variáveis individuais estão variando na mesma direção, o que é esperado para o teste de *α de Cronbach*. No caso de existirem correlações negativas, é preciso multiplicar por – 1 os valores da variável que apresenta al-

terações na direção oposta. A literatura indica como aceitáveis coeficientes entre 0,500 e 0,699 e coeficientes de confiabilidade bons os acima de 0,700. Qualquer resultado de *α de Cronbach* abaixo de 0,500 deve ser interpretado como um indicador formado por variáveis que não representam o mesmo fenômeno ou que não contribuem de maneira coletiva para a explicação de determinadas características do objeto analisado. Nestes casos, recomenda-se uma reformulação dos componentes, excluindo alguns e incluindo outros. Um coeficiente baixo não é necessariamente ruim, pois ele indica que não ocorre na prática o que se esperava de integração entre diferentes conceitos. Em outras palavras, o *α de Cronbach* é capaz de mostrar se as relações semânticas que em teoria fazem sentido são consistentes na prática.

No nosso exemplo, espera-se que um candidato de oposição faça apelo a mudanças, ataque a administração em curso e ataque adversários. No entanto, é possível que na prática esses conteúdos não estejam integrados como se pressupõe a teoria. Então, o *α de Cronbach* irá verificar se quando há presença de uma característica, também tende a existir as outras. No volume I do manual, explicamos como usar o *RCommander*, uma interface do R para testes estatísticos. Por isso, agora vamos direto para os resultados, como segue no quadro abaixo. Seguindo o padrão estabelecido no volume I deste manual, todas as caixas de resultados de testes são divididas em duas partes. Na superior está a linha de comando para que o leitor possa replicar o mesmo teste usando o banco de dados disponível no manual. Em seguida estão os resultados propriamente ditos do teste.

```
Linha de comando - TESTE DE CRONBACH

Rcmdr> reliability
(cov(Dataset[,c("Apelomudanca","AtaqAdmCurs","AtaqAdvers")],
use="complete.obs"))
Resultados – TESTE DE CRONBACH

Alpha reliability =  0.527
Standardized alpha =  0.5769

Reliability deleting each item in turn:
             Alpha Std.Alpha r(item, total)
Apelomudanca 0.6148    0.6306         0.2566
AtaqAdmCurs  0.1840    0.1866         0.5551
AtaqAdvers   0.5067    0.5443         0.2892
```

O teste nos oferece dois coeficientes: o “α” e o “α padronizado”. Isso porque é possível que as variáveis testadas apresentem números distintos de categorias, por exemplo, uma variável pode ter duas, outra pode ter cinco categorias. O resultado para os valores originais é dado pelo primeiro α. Em seguida, há uma padronização pela variável com o menor número de categorias e disso resulta o α padronizado. A comparação entre os dois coeficientes permite que o pesquisador tome a decisão de reduzir ou não o número de categorias nas variáveis originais que vão compor o índice. No nosso caso, como são todas variáveis binárias, os valores dos dois coeficientes ficaram muito próximos entre si (0,527 e 0,577). Nosso coeficiente não é tão alto como esperávamos, mas o menor α ficou acima de 0,500, portanto, podemos dar continuidade ao índice. Vale lembrar que é a decisão final é do pesquisador para continuar ou não com a produção da nova variável e não do *software*. Caberá ao pesquisador justificar o índice, independentemente do coeficiente de confiabilidade.

Uma segunda informação importante que o teste *α de Cronbach* nos dá é o grau de confiabilidade do índice quando se exclui uma das variáveis testadas. Com isso é possível comparar as contribuições individuais de cada variável para o índice final. No exemplo acima, se excluirmos a variável “Apelo a mudanças” o α padronizado sobe para 0,630, enquanto que se excluirmos a variável “Ataque à administração em curso” o α padronizado cai para 0,186. Ou seja, a variável “Ataque à administração” contribui muito mais positivamente para o índice do que a variável “Apelo a mudanças”. Esta última, se excluída, permitiria um incremento no coeficiente final. Esses resultados individuais a partir da exclusão de variáveis podem ser usados para discutir a validade conceitual de determinada variável na composição de um índice. No nosso caso, candidato opositor faz mais ataque à administração em curso do que apelos às mudanças.

Aceitando a confiabilidade no nosso exemplo, dá-se continuidade à criação do indicador aditivo, pela soma de valores das variáveis. Aqui todas têm valores zero ou um para presença ou ausência. Ou seja, a amplitude máxima de variação desse indicador é de três pontos, diferença entre os valores um e três depois da soma, já que consideraremos o zero como ausência de oposicionismo. Aplicando o índice para o banco de dados do HGPE presidencial de 2014, temos o seguinte resultado.

Tabela 3.1. Índice de oposicionismo no HGPE para presidente de 2014

Índice de oposicionismo	Num. de segmentos	Perc. total	Perc. Válido
Não Oposição	506	64,3	
Baixa Oposição	210	26,7	74,7
Média Oposição	35	4,4	12,4
Alta Oposição	36	4,6	12,9
Total	787	100	100

Fonte: autor a partir de dados CPOP

Considerando todos os segmentos do HGPE de 2014, o índice de oposicionismo indica que em 64,3% dos casos não houve presença de nenhum dos fatores que indicam oposição. Assim, o índice inclui 35,7% dos segmentos. Destes, em 74,7% houve baixa oposição, enquanto as distribuições de média e alta ficaram em torno de 12% cada uma. Ou seja, no HGPE dos candidatos não predominou o tom de ataques oposicionistas. Uma vez feito o teste de confiabilidade, aceitado os resultados como válidos e construído o índice pela adição de valores, é possível utilizá-lo em cruzamentos com outras variáveis para explorar o comportamento dos candidatos no HGPE. Por exemplo, poderíamos verificar o comportamento dos três principais concorrentes à presidência em 2014 quando há presença de oposicionismo em seus horários eleitorais. A tabela 3.2 a seguir mostra o número de segmentos no horário eleitoral de Dilma Rousseff, Marina Silva e Aécio Neves. Ela também indica o percentual de segmentos em cada categoria do índice e o percentual válido apenas para o total de segmentos em que houve alguma oposição ao governo em disputa. Para permitir comparações entre os três candidatos, a tabela 3.2 apresenta apenas os resultados para o primeiro turno de 2014. Visto que Marina Silva não disputou o segundo turno, o número de segmentos dela ficaria muito abaixo do apresentado pelos outros dois concorrentes.

Tabela 3.2. Candidato e índice de oposicionismo no HGPE para 1º turno de 2014

Índice de oposicionismo	Dilma Rousseff			Marina Silva			Aécio Neves		
	N.	% Tot.	% Vál.	N.	% Tot.	% Vál.	N.	% Tot.	% Vál.
Não Oposição	224	84,5		27	52,9		75	54	
Baixa Oposição	40	15,1	97,5	14	27,5	58,4	46	33,1	71,8
Média Oposição	1	0,4	2,5	5	9,8	20,8	5	3,6	7,8
Alta Oposição	0	0	0	5	9,8	20,8	13	9,4	20,4
Total	265	100	100	51	100	100	139	100	100

Fonte: autor

A primeira informação que a tabela nos dá é que mesmo no HGPE da candidata à reeleição houve presença de segmentos oposicionistas. Em 15% do total de segmentos de Dilma Rousseff nota-se presença de oposição ao governo em disputa, ainda que em 97,5% deles tenha sido de baixa oposição. Isso é explicado pela presença da variável "ataque a adversários" no índice. Já entre os dois oposicionistas, Marina e Aécio, a distribuição dos percentuais de segmentos com características de oposição foi muito parecida. Em Marina, houve 47,1% de segmentos de oposição (somando as três categorias) e em Aécio o percentual de oposicionismo ficou em 46%. A diferença entre eles está na intensidade. Enquanto Marina teve 58,4% de baixa oposição, Aécio apresentou 71,8% nessa categoria. Como o percentual de média oposição é maior em Marina (20,8%) do que em Aécio (7,8%), podemos identificar que Marina fez oposição de forma mais intensa do que Aécio Neves no HGPE do primeiro turno de 2014.

Neste primeiro tópico do capítulo discutimos, como produzir uma variável categórica pela adição de duas ou mais variáveis. Também vimos como testar a confiabilidade de um índice usando o coeficiente *α de Cronbach* no *RCommander*. Em seguida, apenas a título de exemplo, demonstramos a aplicação de uma variável categórica, o índice de oposicionismo, aplicada ao horário eleitoral da disputa para a presidência do Brasil em 2014. Com isso concluímos a parte inicial que é produzir cruzamentos e análises exploratórias bivariadas, ou seja, usando uma variável para analisar o comportamento de outra. No próximo tópico discutiremos formas de utilizar mais de uma variável em análises de comportamentos. São os chamados testes multivariados. Veremos os seus princípios, os objetivos a serem cumpridos em análises multivariadas, suas limitações e possíveis aplicações à pesquisa em ciência política.

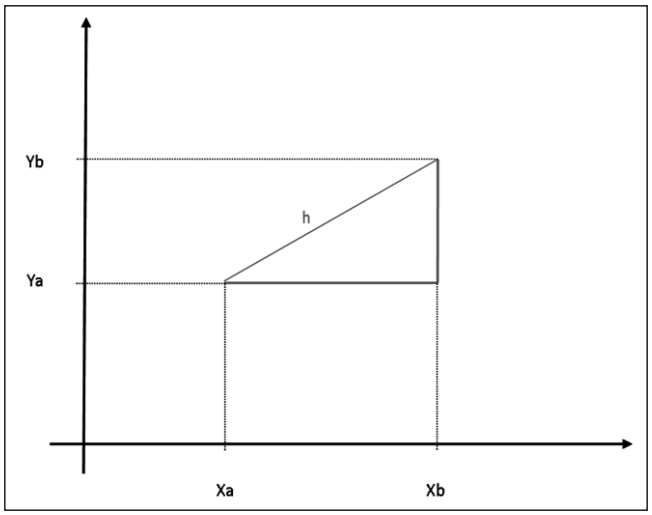
3.3 Testes estatísticos para associações bi e multivariados

A análise multivariada reúne um grande conjunto de técnicas estatísticas que a rigor permite qualquer análise que considere o comportamento de diferentes variá-

veis ao mesmo tempo. As principais técnicas utilizadas e que serão apresentadas a seguir são análise de correspondência simples (*correspondence analysis*), uma variante da análise de correspondência que é a análise de componentes principais, e a análise de agrupamento (*cluster analysis*). Existem outras técnicas, com aplicações variadas, que não serão tratadas aqui. Tanto a análise de agrupamento quanto a de correspondência são duas técnicas ligadas ao conceito de proximidade geométrica, por isso podem ser representadas graficamente. Segundo Pereira (2004), deve-se usar a análise de *cluster* quando se busca identificar grupos de características semelhantes e a análise de correspondência quando se pretende examinar as relações entre categorias de variáveis nominais ou que possam ser tratadas como tal.

Parte-se da ideia de que a relação entre duas variáveis pode ser plotada em um gráfico de coordenadas (x, y) para identificar a localização de cada ponto (A e B) no espaço das variáveis. Uma vez identificados os pontos, é possível traçar uma linha que fará a projeção entre eles. A partir da ligação entre os pontos (Ax e Bx) e (Ay e By) forma-se um triângulo retângulo no qual a distância entre A e B é a sua hipotenusa (h). Outra característica é que esse princípio pode ser replicado várias vezes, permitindo a comparação entre áreas de diferentes variáveis. Sendo assim, a distância entre os dois pontos pode ser calculada pelo teorema de Pitágoras, como indicado no gráfico 3.1 a seguir. Além disso, as relações entre essas distâncias permitem as associações de múltiplas variáveis na descrição de fenômenos ou na explicação de como determinadas características se aproximam ou se distanciam entre si ao mesmo tempo, controladas umas pelas outras. No gráfico, estão representadas posições para duas variáveis (X, Y) de duas observações distintas (a, b). A interseção dos pontos Ya, Xa e Yb, Xb forma uma hipotenusa que tem uma medida. Essa medida de distância permite identificar proximidades espaciais entre as distintas variáveis das observações “a” e “b”.

Gráfico 3.1. Exemplo de gráfico de coordenadas para variáveis A e B



Como a referência espacial está vinculada à geometria plana de Euclides, a distância entre dois pontos calculada dessa forma é chamada de distância euclidiana. Ela é medida em uma unidade comum, ou abstrata, pois não será nem X, nem Y, o que vale tanto para um espaço bidimensional quanto para multidimensional (com vários eixos), já que a distância entre dois pontos será sempre linear e possível de visualização em um plano. Graças a essa característica é que podemos fazer testes de proximidade ou de correspondência entre diferentes pontos distribuídos em planos dimensionais. A primeira técnica que veremos aqui a partir das distâncias euclidianas é a chamada análise de correspondência.

3.3.1 Análise de Correspondência Canônica (ACC)

A análise de correspondência é uma técnica que permite a inclusão de variáveis categóricas nominais ou ordinais com três ou mais categorias no modelo, porém, todas as categorias serão tratadas como variáveis nominais, sem nenhuma hierarquização. Ou seja, não é possível usar para variáveis binárias, pois o cálculo usa categoria como referência e precisa de pelo menos outras duas (dois graus de liberdade) para medir as distâncias. O objetivo é verificar as distâncias entre as categorias, independentemente da ordenação prévia que exista – por isso todas variáveis são tratadas como sendo nominais.

As primeiras noções de análise de correspondência são do início da década de 1930, sendo formalizadas mesmo em 1960 por Benzécri para estudos na área de linguística. Um dos autores que atualizaram os princípios da técnica em estudos mais recentes foi Jobson (1960), que apresenta a técnica como um método de análise estatística multivariada.

É uma das técnicas próprias para análise de variáveis categóricas que tem por objetivo verificar a associação entre as variáveis e suas categorias. Ela permite visualizar na forma de gráfico os resultados de uma tabela de contingência (Mingoti, 2013). Nessas análises, a existência de pelo menos três categorias em cada variável garante que ao utilizar uma delas como referência, outras duas fiquem livres para terem as variações medidas, que é o equivalente ao número mínimo de dimensões do modelo. A análise de correspondência é muito utilizada para examinar relações geométricas de cruzamento e de contingenciamento de variáveis qualitativas. Seu principal objetivo é descrever as relações entre variáveis categóricas nominais em uma tabela de correspondência com baixa dimensionalidade e ao mesmo tempo explorar as relações entre cada categoria das variáveis.

Como ela analisa a distribuição da massa de um conjunto de observações, é uma forma alternativa para verificar o comportamento de dados em uma tabela de contingência. Aqui interessa saber se a massa total das observações está uniformemente distribuída e não se as frequências têm distribuição uniforme, como se dá na análise de resíduos de uma tabela de contingência. No teste, “massa” equivale às frequências marginais da tabela de contingência entre duas variáveis categóricas. A forma como a massa se distribui indica o peso dado às categorias segundo o perfil de distribuição de frequências. Ou seja, o ponto no espaço que apresenta mais “massa” é aquele que atrai mais valores, pois ele representa a(s) marginal(is) com maior frequência de casos. A distância entre cada ponto no gráfico mostra as relações entre as categorias, quanto mais próximos os pontos, maior a correspondência entre eles. Ou seja, maior a tendência de estarem associados.

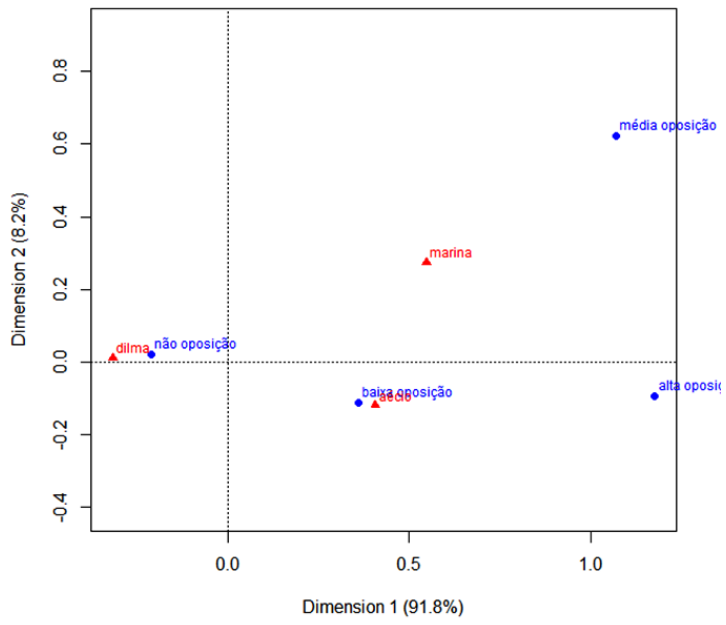
A análise de correspondência é feita a partir da verificação da distribuição das massas em linhas, em colunas ou em ambas simultaneamente. A caracterização de uma linha de frequências segundo a distribuição proporcional das colunas é designada de Perfil de Linha (*row profile*) e o Perfil de Coluna (*column profile*) pode ser obtido da mesma forma. Outra informação importante sobre o teste de correspondência é a média dos perfis, que recebe o nome de Ponto Centroide e representa as frequências marginais

relativas. O teste oferece ainda o resultado de χ^2, que indica o grau de independência de variações da estrutura como um todo. Assim, vale lembrar que ao fazer uma análise de correspondência estamos representando graficamente os resultados que seriam obtidos em uma tabela de cruzamentos, inclusive com o teste de independência de variações. Resumindo, as informações importantes de uma análise de correspondência partem da massa da distribuição, com valores de distâncias em linhas e colunas, um ponto centroide que representa a área mais próxima de todas as frequências marginais relativas e o coeficiente χ^2 para indicar a independência das variações da estrutura de relações.

Para exemplificar uma análise de correspondência, usaremos os dados da tabela 3.2 que cruza os candidatos com as categorias do índice de oposicionismo. Como a interface do *RCommander* não possui a função de análise de correspondência, usarei aqui o teste direto no *RStudio* e apresentarei apenas os resultados. Para mais detalhes sobre como gerar as análises, sugiro consultar a seção de ajuda do *RStudio* para o pacote de análise de correspondência ("ca"). O passo a passo pode ser acompanhado no Anexo 3.6.1. No *RStudio* são necessárias duas operações. Uma para gerar a análise de correspondências e outra para gerar o gráfico de correspondência (Freijo, 2013).

A tabela de dados foi nomeada "ind", de índice no *RStudio*. O comando para a análise de correspondência no *RStudio* é "ca(ind)". O primeiro resultado é de que o modelo foi reduzido a apenas duas dimensões e a primeira dimensão consegue explicar 91,78% das variações, portanto, trata-se da principal dimensão do modelo. Em seguida, o *software* oferece as estatísticas para as categorias nas linhas e para as colunas. São elas a massa, o qui-quadrado da distância e a inércia. Também oferece as coordenadas de cada categoria para as dimensões 1 e 2, caso o pesquisador queira copiar os valores em um editor de gráficos e produzir seu próprio gráfico. Percebe-se que a maior massa é de "não oposição", seguida de "oposição baixa", como indicado na tabela de cruzamentos. São as categorias com maior número de segmentos de horário eleitoral. Em seguida aparecem as mesmas estatísticas para as categorias das colunas, no caso, os candidatos. A maior massa é para Dilma, seguida de Aécio, novamente, indicando as maiores concentrações de casos. O teste de χ^2 para as duas variáveis mostrou-se altamente significativo, com coeficiente de 70,918 para 6 graus de liberdade, o que indica significância estatística para dependência das variações de categorias das duas variáveis.

```
Banco de dados: ind
                dilma marina aécio
não oposição      224     27    75
baixa oposição     40     14    46
média oposição      1      5     5
alta oposição       0      5    13

Linha de comando: ca (ind)
Resultado: Análise de Correspondência

Principal inertias (eigenvalues):
           1        2
Value      0.143052 0.012812
Percentage 91.78%   8.22%


 Rows:
        não oposição baixa oposição média oposição alta oposição
Mass        0.716484       0.219780       0.024176      0.039560
ChiDist     0.212341       0.377130       1.238470      1.181442
Inertia     0.032305       0.031259       0.037081      0.055219
Dim. 1     -0.559232       0.951511       2.828289      3.113749
Dim. 2      0.165414      -0.996091       5.513746     -0.831499


 Columns:
            dilma   marina     aécio
Mass     0.582418 0.112088  0.305495
ChiDist  0.318578 0.612036  0.423403
Inertia  0.059111 0.041987  0.054766
Dim. 1  -0.841868 1.448060  1.073697
Dim. 2   0.090773 2.413439 -1.058563
Chi-squared test
data:  d
X-squared = 70.918, df = 6, p-value = 2.65e-13
> plot (ca(ind))
```



A melhor visualização dos dados se dá a partir do gráfico de correspondência, que é gerado a partir do comando “plot (ca(ind))”. Nele fica evidente a correspondência entre Dilma e “Não Oposição” e entre Aécio e “Baixa Oposição”. Marina fica em um ponto quase equidistante entre as três categorias de oposicionismo. A interseção das duas linhas tracejadas, com origem na posição zero das dimensões 1 e 2, indica o ponto centroide e, em consequência, quais pares de categorias estão mais próximos e mais distantes dele.

O pacote de análise de correspondência no R não oferece diretamente uma tabela de perfis de linhas e colunas. Para produzir uma tabela que sumarize tanto os valores das colunas quanto das linhas, basta usar a massa de cada uma delas como marginais e multiplicar os valores obtidos nas linhas e nas colunas para cada casa. Assim se tem o produto da categoria da linha com a da coluna, conforme apresentado na tabela 3.3 a seguir:

Tabela 3.3. Produto das massas das linhas e colunas da análise de correspondência

Índice	Dilma	Marina	Aécio	Massa linhas
Não oposição	**0,416**	0,080	**0,218**	0,716
Baixa oposição	0,127	**0,024**	0,066	0,219
Média oposição	0,013	0,002	0,007	0,024
Alta oposição	0,022	0,004	0,011	0,039
Massa colunas	0,582	0,112	0,305	1,000

Fonte: autor

A informação nova que a tabela de produtos das massas nos traz é que as maiores concentrações dos candidatos, quando consideradas as correspondências totais se dá entre Dilma e “Não oposição” e Aécio e “Não oposição”. Marina está distribuída de forma mais homogênea, o que dificulta correspondências. Ainda assim, sua maior massa está em “Baixa oposição”. Em resumo, a análise de correspondência nos permite verificar as associações entre pares de categorias, controlando-as pelas categorias de outras variáveis.

A análise de correspondência é uma importante técnica que auxilia na análise de tabelas de contingência, em especial quando há um grande número de categorias envolvidas. Impressiona o número de informações que se pode extrair de um gráfico de correspondência como no exemplo acima. Vale lembrar que usamos uma variável

ordinal (o índice de oposicionismo) em um teste que desconsidera a transitividade entre categorias, portanto, os resultados devem ser lidos como entre duas variáveis nominais. No teste de correspondência o mais importante é verificar o ajuste do modelo para análise a partir dos valores de inércia e percentual de variações explicadas por dimensão, além de reconhecer as relações entre as variáveis e categorias através de sua proximidade no gráfico plano.

No entanto, a análise de correspondência canônica discutida até aqui apresenta três grandes limitações. A primeira diz respeito ao número de variáveis no modelo, que deve se limitar a duas. A segunda é que ela só permite trabalhar com variáveis categóricas nominais. Por fim, as variáveis nominais precisam ter três ou mais categorias. Para o caso de existir a necessidade de incluir mais de duas variáveis no modelo e/ou elas serem ordinais, deve-se usar o teste dos componentes principais. Se o teste for entre três ou mais variáveis categóricas nominais, o indicado é que sejam feitas as múltiplas correspondências. Se for entre mais de duas variáveis ordinais, deve-se usar o teste de componentes principais. É o que apresentaremos a seguir.

3.3.2 Teste de Múltipla Correspondência

No caso da análise de múltipla correspondência, medem-se as presenças de categorias como se todas fossem nominais, agrupando-as de forma que os casos mais próximos são aqueles que apresentam valores parecidos. Assim, as categorias terminam sendo divididas em subgrupos homogêneos. Quanto mais próximos forem os códigos das categorias das variáveis, mais homogêneos serão os subgrupos. Aqui, o princípio é o mesmo que o da ACC, porém utilizando três ou mais variáveis ao mesmo tempo.

O exemplo está dividido em duas partes. Na primeira, são apresentados os resultados para o banco de dados de todos os programas de HGPE do primeiro turno de 2014 para presidente. Em seguida, o banco é dividido pelos principais candidatos, para compararmos as diferentes correspondências entre as variáveis. De início, foram usadas cinco variáveis na análise de correspondência múltipla. A unidade de análise é o candidato e dia de exibição para o banco completo e apenas o dia de exibição para os

bancos de cada candidato. As variáveis de início no modelo foram:

- Apelo à mudança: se o candidato fez apelo direto à mudança da situação do País no programa;

- Índice de Oposicionismo: mede a intensidade de oposição ao governo e aos demais concorrentes em cada programa;

- Proporção de tempo: em faixas de percentuais, para diferenciar programas que tiveram segmentos mais curtos de programas com segmentos mais longos.

- Ofensiva quanto a temas: para quando o candidato usa o espaço do horário eleitoral para apresentar propostas de projetos de políticas públicas; e

- Associação à administração em curso: identifica se o candidato se associa, ainda que indiretamente, a algum evento ou parte do governo em disputa.

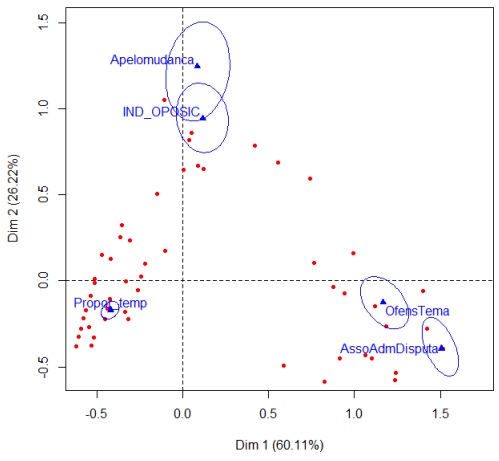
A hipótese é que os candidatos separam os espaços do horário eleitoral em uma parte para tratar de temas estritamente políticos (fazer oposição, pedir mudança, etc.) e outra parte para apresentar propostas de políticas públicas e tratar de temas propriamente ditos. A correspondência aqui fica por conta do tempo usado pelos candidatos para cada um desses conjuntos de variáveis.

O teste de múltipla correspondência foi rodado no *RCommander* e a saída de resultados é apresentada no quadro abaixo. O coeficiente para todo o teste é o $\chi 2$, que para todo o banco foi estatisticamente significativo (1.442,17). Em seguida, estão plotados os *eigenvalues* das variâncias explicadas por dimensão. A primeira dimensão é a mais explicativa, com explicação de 0,456 de variância, o que representa um percentual de 60,11%. Se somada à segunda dimensão, a variância total explicada por elas é de 86,33%. Quanto aos valores individuais das variáveis no modelo de correspondência, tanto os coeficientes das colunas quanto o gráfico logo a seguir, mostram dois pares de variáveis em correspondência e a última isolada em um quadrante. Trata-se de Apelo à mudança e oposicionismo por um lado e ofensiva quanto a tema e associação à administração em disputa por outro.

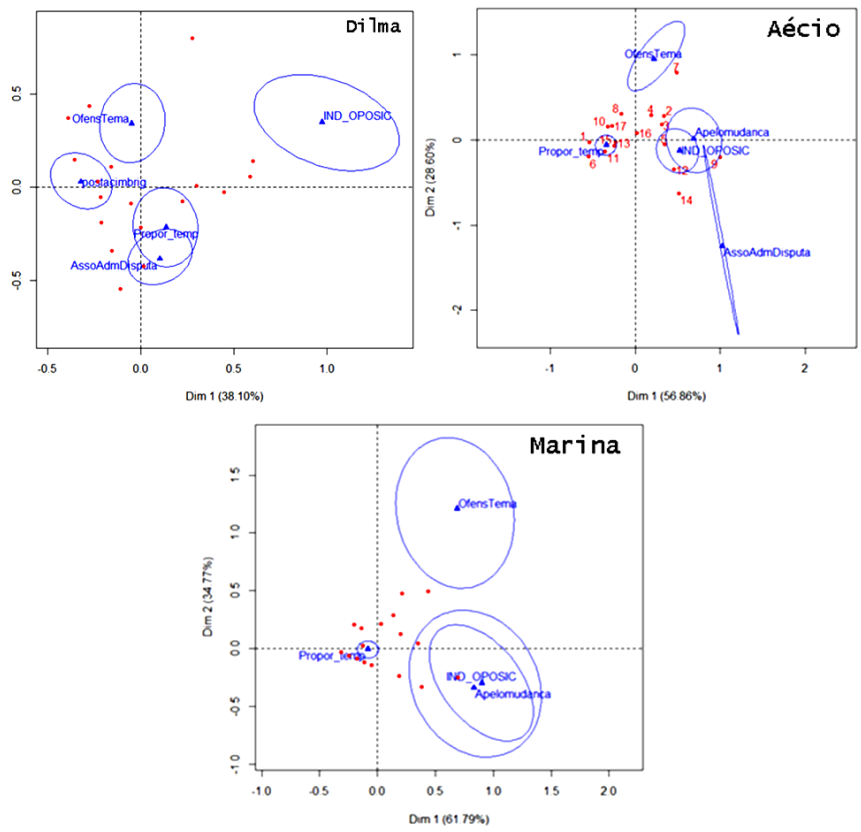
Como esperado, no HGPE, quando se usa o tempo para fazer oposição ao governo em disputa, isso corresponde ao tempo destinado aos apelos à mudança. Por outro lado, quem mais usa o tempo para falar de propostas e temas é quem também se associa à administração em curso. A proporção de tempo está distante de todas as demais variá-

veis, indicando que tanto segmentos grandes quanto pequenos se distribuem de maneira homogênea entre todas elas. Além dos coeficientes para as variáveis (colunas), o teste de múltipla correspondência também oferece informações para os casos (linha). No nosso exemplo, cada programa é um caso. E estão representados no gráfico de múltipla correspondência por pontos vermelhos. Isso nos permite perceber que existem alguns programas onde há grande presença de conteúdo oposicionista, para os pontos que estão dentro do círculo desta variável. E assim sucessivamente. Os pontos que se encontram fora do espaço de abrangência de todas as variáveis representam programas nos quais não houve predomínio de nenhuma das características inseridas no modelo de análise.

```
Linha de comando:
Rcmdr>  res<-CA(hgpe.CA, ncp=5, row.sup=NULL, col.sup=NULL, graph = FALSE)
Rcmdr>  ellipseCA(res, ellipse=c("col"), axes=c(1, 2), col.row="red",
Rcmdr+    col.col="blue", label=c("col", "col.sup", "row.sup"), title="")
Rcmdr>  summary(res, nb.dec = 3, nbelements=10, nbind = 10, ncp = 3, file="")

Resultados:
Call:"res<-CA(hgpe.CA, ncp=5, row.sup=NULL, col.sup=NULL, graph = FALSE)"

The chi square of independence is equal to 1442.17 (p-value = 9.112937e-182 ).

Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4
Variance               0.456   0.199   0.090   0.013
% of var.             60.113  26.217  11.924   1.746
Cumulative % of var.  60.113  86.330  98.254 100.000

Columns
                Iner*1000     Dim.1      ctr    cos2     Dim.2      ctr    cos2
Propor_temp     |   133.179 |  -0.419   25.133   0.860 |  -0.169    9.367   0.140
AssoAdmDisputa  |   225.742 |   1.507   37.189   0.751 |  -0.393    5.812   0.051
Apelomudanca    |    92.380 |   0.087    0.090   0.004 |   1.244   42.564   0.916
OfensTema       |   216.398 |   1.168   37.308   0.786 |  -0.124    0.966   0.009
IND_OPOSIC      |    90.347 |   0.117    0.280   0.014 |   0.942   41.291   0.908
```



```
Gráficos de múltipla correspondência para cada um dos principais
candidatos:
```



A última parte do quadro apresenta três gráficos de múltipla correspondência, um para cada um dos principais candidatos a presidente em 2014. O objetivo é mostrar as diferenças de comportamentos das variáveis entre os candidatos. Foram mantidas as mesmas variáveis iniciais para todos eles, porém no caso de Marina Silva foi excluída a associação à administração por não apresentar valores válidos. Os principais resultados para Dilma são que a proporção de tempo coincide no espaço com associação à administração em disputa, o que indica que nos programas dela os segmentos destinados a falar do governo tenderam a ser mais longos que os demais. Ofensiva em relação a tema está distante, o que significa que as proposta foram apresentadas em momentos distintos daqueles em que se falava do governo. No caso de Aécio, o gráfico mostra um comportamento de oposicionista, que associa fortemente a presença do índice de oposição ao apelo à mudança. Associação à administração forma uma área curiosa, pois ela não tem alta frequência, por isso o isolamento em um quadrante. Mas quando

aparece, vincula-se ao discurso de oposição e de apelo à mudança. Por fim, Marina Silva apresenta o gráfico de correspondência mais próximo do esperado para uma candidata de oposição. Os programas que tratam dos temas político ocupam praticamente o mesmo espaço, sobrepondo apelo à mudança ao índice de oposicionismo. Em outro espaço, encontram-se as propostas de políticas para temas públicos.

Até aqui, as análises de correspondência, canônica e múltipla, foram úteis para identificar relações entre variáveis, entre casos e entre variáveis e casos a partir da plotagem em um espaço bidimensional e pelas distâncias euclidianas. O próximo teste tem um objetivo distinto. Chamado de Análise de Componentes Principais, ele serve para extrair das variáveis vetores que não estão correlacionados entre si. Sua característica mais importante para nós é a de redução do número de variáveis na análise e interpretação a partir das combinações lineares produzidas (Mingoti, 2013).

3.3.3 Análise de Componentes Principais (PCA)

A análise de componentes principais é um método usado para a redução de dimensões entre variáveis categóricas ordinais, no qual também podem ser usadas variáveis contínuas. As variáveis são analisadas em conjunto para identificar as dimensões principais de variação. No caso das variáveis contínuas, o teste cria categorias para deixar a distribuição o mais uniforme possível. A partir desse método é possível encontrar relações entre as variáveis, entre as categorias e entre variáveis e categorias. O teste quantifica os casos observados e as categorias das variáveis para medir o grau de associação entre elas.

No exemplo para Análise de Componentes Principais, continuaremos usando o banco de dados do HGPE de 2014. Para rodar o teste do *RCommander* é preciso instalar um plug-in na interface chamado de "*FactoMiner*". As instruções para instalação encontram-se no CRAN do R. (Ver o passo a passo da instalação no anexo 3.6.2). Para interpretar os resultados da análise de componentes principais é preciso saber o que representam e qual a direção das categorias de cada variável. No nosso caso, usaremos cinco variáveis na análise. Elas estão compostas da seguinte forma, do menor para o maior valor:

- Semana – semana da campanha em que o HGPE foi ao ar, vai de 1 até 15.

- Tema_Aggr – tema agregado, reúne os temas em três grandes conjuntos: 1-tema público propriamente dito; 2-formação da imagem do candidato; 3-metacampanha.

- Candidato – codificado pelo número do partido do candidato: 13-Dilma; 40-Marina; 45-Aécio;

- Ind_oposic – índice de oposicionismo, como já apresentado, possui quatro categorias ordinais: 0-sem oposição; 1-baixa oposição; 2-média oposição; 3-alta oposição.

- Prop_Cat – proporção da participação do segmento no programa do candidato em categorias ordinais. A variável foi organizada em três categorias: 1-terço inferior; 2-terço médio; 3-terço superior em duração de segmentos.

Conhecer os códigos das variáveis e as direções é fundamental para interpretar os resultados dos componentes principais. Aqui, os valores numéricos das categorias serão comparados conjuntamente e o resultado mostra quais são as ocorrências mais comuns entre todos os casos. Quanto mais próximas as categorias de diferentes variáveis, mais elas se apresentam em conjunto nos casos. O *output* a seguir reproduz os principais coeficientes produzidos no *RCommander* para o teste de componentes principais.

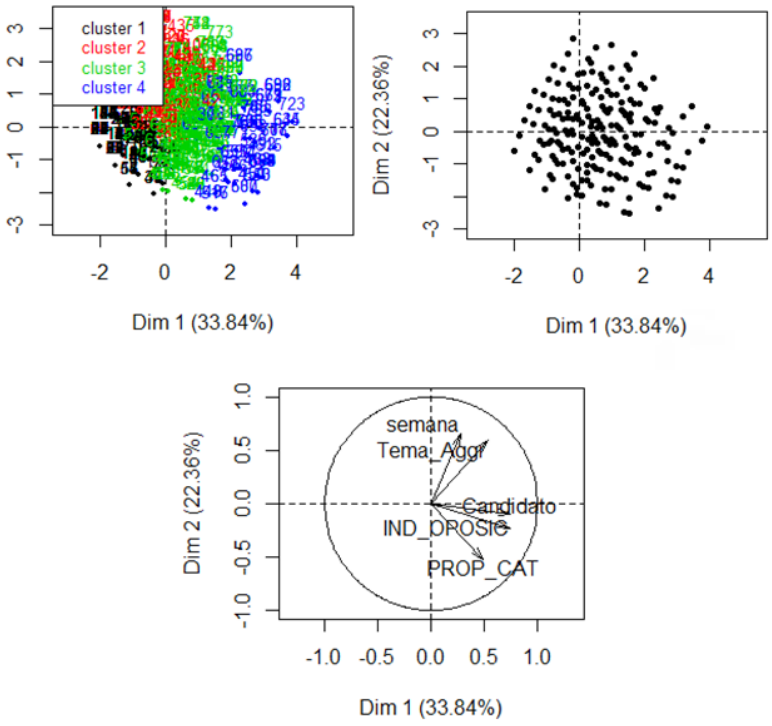
Os resultados começam com o percentual de variação explicado em cada dimensão e as variâncias (*eigenvalues*). O modelo gerou cinco dimensões. Percebe-se que a dimensão 1 explica 33,8% das variâncias e a dimensão 2 explica 22,3%. Somadas, elas resultam em 56,1% do total da variância. Não é tão alta como se espera em análises de componentes principais para as duas primeiras dimensões. Isso significa que as categorias não se distribuem na forma de componentes específicos. Elas tendem a ser mais dispersa. A segunda informação importante no teste é sobre a posição de cada variável nas dimensões. Quanto mais próximos os valores em uma dimensão, maior a possibilidade das categorias das variáveis estarem formando um componente. No caso, na dimensão 1 todos os valores são positivos, indicando que as variáveis vão para a mesma direção nessa dimensão, ainda que índice de oposicionismo e candidato apresentem coeficientes mais próximos entre si, acima de 0,700. O que permite a diferenciação é a dimensão 2. Nela, semana e tema agregado apresentam coeficientes positivos, indicando uma direção, e as outras três variáveis têm coeficientes negativos, que aponta para outra direção. Os coeficientes que seguem no modelo apenas reforçam os componentes apresentados nas dimensões, pois o valor de contribuição de cada

variável e as correlações nas dimensões mostram que as categorias de semana e tema agregado formam um componente e as outras três variáveis formam outro componente.

O componente formado por semana e tema agregado nos informa que, conforme a campanha se aproxima do final, os candidatos tendem a ocupar mais seus horários eleitorais com conteúdos de metacampanha, que é falar sobre os eventos da própria campanha. Isso porque a variável semana apresenta códigos crescentes de 1 a 15 e na variável tema agregado metacampanha é o código mais alto. Assim, conforme cresce semana, também aumenta a presença de metacampanha. No segundo componente da dimensão 1, se olharmos com atenção, perceberemos uma proximidade maior entre candidato e índice de oposicionismo, com a proporção de tempo um pouco mais distante. Isso significa que nesse componente estão mais associadas as categorias das duas primeiras variáveis. Assim, como sabemos que na variável candidato Dilma é o código mais baixo e Aécio o mais alto, fica evidente que o índice de oposicionismo cresce, conforme aumenta o código do candidato. Em outras palavras, Dilma apresenta oposicionismo mais baixo, como já demonstrado nos exemplos anteriores.

```
Linha de Comando:
Rcmdr>  Dataset.PCA<-Dataset[, c("semana", "Candidato", "PROP_CAT", "Tema_Aggr",
Rcmdr+    "IND_OPOSIC")]
Rcmdr>  res<-PCA(Dataset.PCA , scale.unit=TRUE, ncp=5, graph = FALSE)
Rcmdr>  res.hcpc<-HCPC(res ,nb.clust=-1,consol=TRUE,min=3,max=10,graph=TRUE)
Rcmdr>plot.PCA(res,axes=c(1,2),choix="ind",habillage="none",col.ind="black",
Rcmdr+ col.ind.sup="blue",col.quali="magenta",label=c("ind.sup","quali"),
Rcmdr+    new.plot=TRUE, title="")
Rcmdr>plot.PCA(res,axes=c(1,2),choix="var",new.plot=TRUE,col.var="black",
Rcmdr+col.quanti.sup="blue",label=c("var","quanti.sup"), lim.cos2.var=0,
Rcmdr+    title="")
Rcmdr>  summary(res, nb.dec=3, nbelements=10, nbind=10, ncp=3, file="")
Call:
"res<-PCA(Dataset.PCA , scale.unit=TRUE, ncp=5, graph = FALSE)"
Principais Estatística:
Eigenvalues
                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
Variance                1.692   1.118   0.915   0.671   0.605
% of var.              33.838  22.360  18.292  13.416  12.094
Cumulative % of var.   33.838  56.198  74.490  87.906 100.000
Variables
             Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
semana     |  0.280  4.647  0.079 |  0.657 38.586  0.431 |  0.648 45.947  0.420
Candidato  |  0.741 32.476  0.549 | -0.099  0.885  0.010 | -0.332 12.038  0.110
PROP_CAT   |  0.490 14.216  0.241 | -0.523 24.461  0.273 |  0.530 30.664  0.280
Tema_Aggr  |  0.530 16.587  0.281 |  0.593 31.419  0.351 | -0.320 11.217  0.103
IND_OPOSIC |  0.737 32.074  0.543 | -0.228  4.649  0.052 | -0.035  0.135  0.001
```

```
$contrib
               Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
semana       4.64696 38.5860426 45.9467653  6.5548679  4.2653639
Candidato   32.47596  0.8845328 12.0378313  0.5455169 54.0561603
PROP_CAT    14.21598 24.4611706 30.6639682 30.5090476  0.1498295
Tema_Aggr   16.58676 31.4189543 11.2167376 28.6498407 12.1277076
IND_OPOSIC  32.07434  4.6492997  0.1346976 33.7407268 29.4009387

Rcmdr>  dimdesc(res, axes=1:5)
$Dim.1
$Dim.1$quanti
           correlation       p.value
Candidato    0.7412588 4.705131e-138
IND_OPOSIC   0.7366610 1.686526e-135
Tema_Aggr    0.5297484  3.818497e-58
PROP_CAT     0.4904303  7.316500e-49
semana       0.2803970  1.101835e-15

$Dim.2
$Dim.2$quanti
            correlation       p.value
semana       0.65680491 2.512391e-98
Tema_Aggr    0.59267572 8.262132e-76
Candidato   -0.09944392 5.234423e-03
IND_OPOSIC  -0.22798960 9.725528e-11
PROP_CAT    -0.52294967 1.883568e-56
```



Além dos coeficientes, a análise de componentes principais produz três gráficos apresentados acima. O primeiro deles indica a distribuição dos *clusters*. Como

indicado nos coeficientes gerais do teste, não houve boa distribuição das variações, o que resulta em *clusters* menos evidentes. De qualquer maneira, são quatro *clusters*, representados pelo número do caso no banco de dados. Apesar da sobreposição de códigos, é possível perceber que no *cluster* de cor preta estão as semanas iniciais e no azul as últimas semanas de HGPE. O segundo gráfico mostra os pontos da distribuição dos casos no espaço bidimensional, indicando a ausência de viés ou tendência clara nas distribuições individuais – que aqui são as semanas.

O gráfico das variáveis é o mais importante aqui, pois ele representa as direções e proximidades entre variáveis nas duas principais dimensões do modelo. Nele fica claro que semana e tema agregado formam um componente, enquanto as outras três variáveis formam outro componente. Comparando com o gráfico dos *clusters* é possível afirmar que conforme se aproxima o fim da campanha há uma tendência de se encontrar segmentos com maior duração, conteúdo predominante de metacampanha e mais oposicionismo, principalmente em Aécio e Marina. Ou, visto de outra forma, no início do horário eleitoral há menos oposicionismo, mais temas públicos (propostas de políticas) e segmentos com menor duração do que no final da campanha. Na análise de componentes principais, os *clusters* são apresentados como acessórios para o resultado. No próximo tópico discutiremos os objetivos e princípios da técnica de análise por agrupamentos.

3.3.4 Análises de agrupamentos (*Cluster*)

Um *cluster* é um grupo relativamente homogêneo de casos ou observações. Nesse tipo de análise exploratória são calculadas as distâncias entre objetos em um espaço multiplano representado por eixos de todas as variáveis. Aqui, a análise pode ser entre categorias ou entre variáveis. No caso de analisar os *clusters* entre variáveis, elas podem ser transformadas em binárias, o que facilitará a interpretação dos resultados. Nesse caso os agrupamentos serão entre as variáveis que apresentam a presença da característica (código=1) contra as que não apresentam a característica (código=0). As variáveis – ou categorias – são agrupadas em função da proximidade mútua. Começa por constituir um grupo inicial os dois objetos (variáveis) mais próximos. Depois, identifica-se qual a variável que se localiza mais próxima do centro desse grupo e forma-se

um novo grupo até o total de objetos estudados. Todos os coeficientes são calculados a partir das distâncias planas euclidianas. Os *clusters* também são reduções dimensionais, ou seja, eles permitem diminuir o número de dimensões da realidade, o que facilita a interpretação, porém, isso também diminui o detalhamento dos fenômenos.

Existem diferentes estatísticas nas análises de *clusters*, mas os gráficos bidimensionais gerados pela proximidade de variáveis, o *cluster* por *kmeans* e o dendrograma no *cluster* hierárquico são os mais informativos. Aqui, demonstraremos a aplicação da análise de *clusters* em ciência política usando novamente o banco de dados do HGPE para presidente de 2014. Vamos gerar *clusters* para cada um dos três principais candidatos na disputa: Marina Silva, Dilma Rousseff e Aécio Neves. Não faria sentido algum rodar uma análise de *cluster* reunindo os dados de todos os três candidatos em um mesmo espaço. A unidade de análise é o programa, portanto, os valores são as somatórias de ocorrências de cada variável em um programa. As variáveis que utilizaremos no teste são as seguintes:

- Ataque a adversários: quantas vezes o candidato atacou outros candidatos no programa;

- Endosso de liderança política: quantas vezes o programa foi usado para que outros políticos endossassem a figura pública do candidato;

- Menção a partido: quantas vezes o candidato mencionou seu próprio partido no programa eleitoral;

- Ofensiva quanto a temas: quantas vezes o candidato apresentou propostas ou projetos relacionados diretamente a temas de políticas públicas.

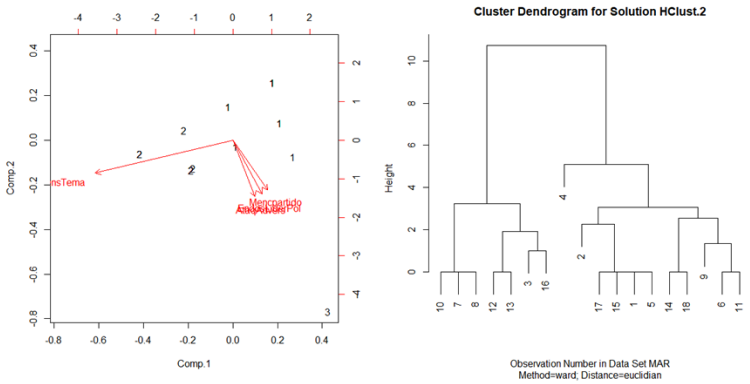
O objetivo da análise exploratória é verificar o comportamento dos *clusters* nos programas de candidato, ou seja, o quanto que essas variáveis se aproximaram ou se distanciaram no HGPE. A hipótese é que os candidatos dividam os programas em momentos para falar de política partidária, logo, espera-se menção a partido e endosso de liderança devem ficar próximos entre si, e em outros momentos os programas tratem de propostas propriamente ditas, quando devem aparecer as ofensivas em relação a temas. A variável ataque a adversários é uma incógnita, pois pode estar tanto próxima de políticas públicas, quando as propostas são acompanhadas de críticas ao que já foi feito, ou pode ser um ataque político e se aproximar de menção a partido. O teste foi rodado no *RCommander* e as saídas dos resultados seguem no quadro abaixo.

Na análise de *clusters* por *kmeans* quem define o número de *clusters* é o pesquisador. Existem muitas formas para definição desse número, mas sempre é bom contar com o bom senso do pesquisador. Aqui, eu escolhi três *clusters* por estarmos trabalhando com quatro variáveis. A análise hierárquica (discutida nos gráficos mais abaixo) já permite a identificação de um número adequado de *clusters* em função das distribuições das variáveis. A primeira informação no resultado do *cluster* é o número de casos em cada um deles. No HGPE da Marina, o maior *cluster* é o primeiro, com 10 casos, seguido de sete e apenas um no terceiro. Isso significa que se deve preferir o primeiro *cluster* para analisar os valores individuais. A segunda informação é o valor que representa o ponto centroide de cada variável por *cluster*. Olhando para a linha do *cluster* 2, os centroides mostram que ataque a adversário (0,42), endosso de líder (0,14) e menção a partido (0,0) ficam próximos entre si, enquanto ofensiva em relação a temas (2,28) é o mais distante no espaço. Quanto é longo e quanto é próximo? Não existe limite. Sempre a análise será relativa. Um valor é próximo ou distante em relação aos demais valores.

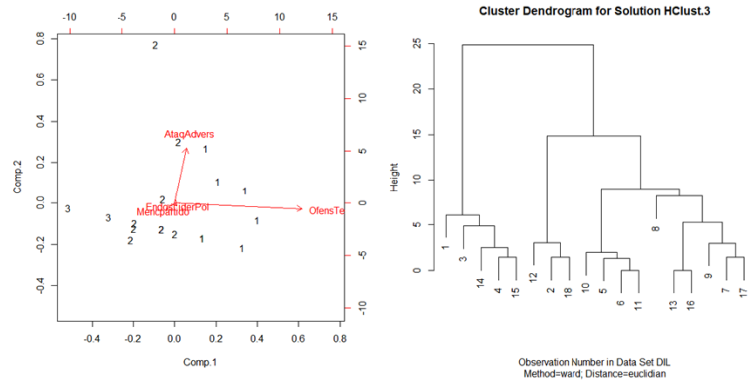
Se para Marina Silva a variável ofensiva em relação a temas já se apresentava como a mais distante das demais, no caso de Dilma Rousseff, essa distância aumenta ainda mais. Já no caso de Aécio, o comportamento é distinto do verificado nas duas candidatas. Há distância da variável endosso de liderança política em um *cluster* e ofensiva quanto a temas em outro. Isso indica que o HGPE de Aécio Neves apresentou formato distinto dos dois anteriores no que diz respeito à participação de endosso de lideranças políticas. Essas diferenças comparativas entre os candidatos podem ser visualizadas nos gráficos de *cluster* e no dendrograma que estão no final do quadro.

```
Marina
Rcmdr>.cluster<- Kmeans(model.matrix(~-1 + AtaqAdvers + EndosLiderPol +
Rcmdr+Mencpartido+OfensTema, MAR), centers=3, iter.max=10, num.seeds=10)
Rcmdr>  .cluster$size # Cluster Sizes
[1] 10  7  1

Rcmdr>  .cluster$centers # Cluster Centroids
  new.x.AtaqAdvers new.x.EndosLiderPol new.x.Mencpartido new.x.OfensTema
1        0.3000000           0.1000000               0.1        0.300000
2        0.4285714           0.1428571               0.0        2.285714
3        2.0000000           2.0000000               2.0        0.000000

Dilma
Rcmdr>.cluster <- Kmeans(model.matrix(~-1 + AtaqAdvers + EndosLiderPol +
Rcmdr+Mencpartido+OfensTema, DIL), centers=3, iter.max=10, num.seeds=10)
Rcmdr>  .cluster$size # Cluster Sizes
[1] 7 9 2
```

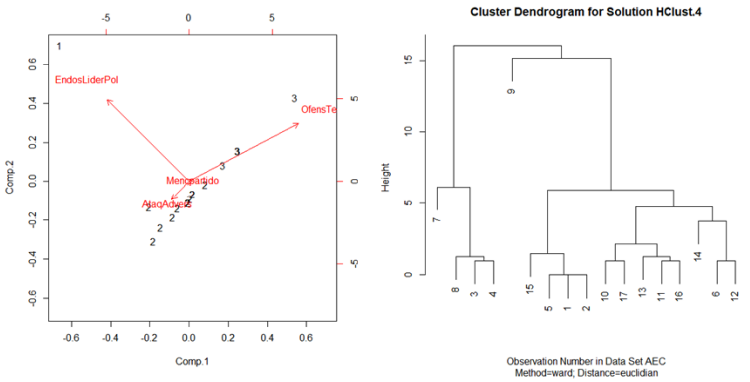
```
Rcmdr>  .cluster$centers # Cluster Centroids
  new.x.AtaqAdvers new.x.EndosLiderPol new.x.Mencpartido new.x.OfensTema
1         1.142857           0.2857143          0.0000000       13.571429
2         1.111111           0.3333333          0.5555556        8.444444
3         0.000000           0.0000000          0.0000000        3.500000

Aécio
Rcmdr>.cluster<- Kmeans(model.matrix(~-1 + AtaqAdvers + EndosLiderPol +
Rcmdr+ Mencpartido+OfensTema,AEC), centers=3, iter.max=10, num.seeds=10)
Rcmdr>  .cluster$size # Cluster Sizes
[1]  1 12  4

Rcmdr>  .cluster$centers # Cluster Centroids
  new.x.AtaqAdvers new.x.EndosLiderPol new.x.Mencpartido new.x.OfensTema
1         1.000000          9.00000000         0.00000000        0.000000
2         1.166667          0.08333333         0.08333333        1.416667
3         0.250000          0.00000000         0.25000000        5.750000
```



Marina



Dilma



Aécio

O gráfico do *cluster* de Marina Silva apresenta a maior distância entre ofensiva em relação a temas e as outras variáveis. Como ofensiva a temas está muito distante, todas as outras se sobrepõem no espaço. Isso indica que, no caso de Marina, os programas eleitorais que tratavam de proposta de temas eram separados dos programas sobre assuntos políticos, apoios partidários ou ataques a adversários. Já nos casos de Dilma e Aécio os programas eleitorais se agrupam em três posições. Em Dilma, há uma posição para ofensiva quanto a temas, outra para ataques a adversários e a terceira formada por menção a partido e endosso de lideranças. Assim como Marina, os programas de Dilma separam a apresentação de propostas de outros assuntos, porém, aqui ela também divide aspectos positivos como endosso de lideranças com menção a partido de um lado e, de outro, o ataque a adversários. Ainda que também se agrupe em três pontos, há uma diferença em relação ao horário eleitoral de Aécio. Assim como as outras, ele separa propostas de políticas públicas dos temas político-partidários, porém, ele agrupa ataque aos adversários com menção ao seu partido, enquanto endosso de lideranças está em outro *cluster*. Daí é possível considerar que a estratégia de Aécio foi atacar seus adversários citando o próprio partido como alternativa ou solução para as críticas.

Por fim, o gráfico dendrograma nos mostra como as variáveis se organizam não mais de forma aleatória, mas seguindo uma hierarquia de maiores semelhanças e dessemelhanças. O eixo Y nos mostra o ponto de referência para o corte que vai resultar em um número de *clusters*. Quanto maior o valor de Y, menor o rigor. Por exemplo, no dendrograma de Marina, se traçar uma reta a partir do número quatro do eixo Y, perceberemos que ele nos dará apenas três *clusters*. Um formado pelas semanas que vão de 10 a 16, outro formado pela semana quatro e o terceiro pelas demais semanas. Agora, se partirmos do ponto quatro nos dendrogramas de Dilma e Aécio, perceberemos que o número de *clusters* formados será bem maior. Isso se deve pelo fato de os programas eleitorais da candidata Marina apresentarem maiores semelhanças do que os dos outros candidatos no que diz respeito às variações das características inseridas no modelo.

Neste capítulo, tomamos conhecimento de alguns dos principais testes com dados categóricos e discretos para relações bi e multivariadas usando pacotes e *plug-ins* disponíveis no *RCommander*. Os testes de correspondência, componentes principais e *cluster* servem para relacionar ou encontrar as forças de associações entre variáveis a

partir de suas unidades originais ou por transformações em vetores matemáticos. Como todo modelo estatístico, os testes discutidos aqui também apresentam limitações. Por isso não podem se transformar em um fim em si mesmo.

Testes servem para nos levar de um lugar para outro em relação a nosso objeto de análise e não nos manter distantes do mundo empírico. Por isso, a recomendação sempre é ir com calma nas análises e modelagens mais complexas. Não pensar duas vezes em mudar de caminho se o meio começar a parecer mais importante que o fim. Nosso objetivo continua sendo usar as ferramentas naquilo que elas podem ser úteis, para a análise em ciência política e não submeter a análise política às limitações dos modelos estatísticos. No próximo capítulo, estudaremos algumas técnicas de análise de redes sociais, na qual a principal característica é a transformação da unidade de análise, que deixa de ser o indivíduo e passa a ser as relações entre indivíduos.

3.4 Referências bibliográficas do Capítulo III

Cronbach, L. J. (1971). Test Validation. In: R. L. Thorndike (Ed.), *Educational Measurement*. Washington: American Council on Education, 443-507.

Freijo, J. B. (2013). El paquete estadístico R. *Cuadernos metodológicos*, 48. Madrid: Centro de Investigaciones Sociológicas.

Lé, S., *et al.* (2008). FactoMiner: An R package for multivariate Analysis. *Journal of Statistical Software*, 25(1), 1-18.

Mingoti, S. A. (2013). *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belo Horizonte: Editora UFMG.

Pereira, J. C. R. (2004). *Análise de Dados Qualitativos – estratégias metodológicas para as Ciências da Saúde, Humanas e Sociais*. São Paulo: EdUsp/Fapesp.

3.5 Exercícios propostos do Capítulo III

3.5.1 A partir do banco de dados **BDCAP3V2_HGPE** disponível em *https://blogempublico.files.wordpress.com/2018/02/bdcap3v2_hgpe.xlsx*, rode um teste de *α de Cronbach* para um índice que chamaremos de índice de governismo. As variáveis fazem parte de um conjunto de dados extraídos das edições diárias do Horário Gratuito de Propaganda Eleitoral, do primeiro turno das eleições de 2014, para o cardo de presidente da república. São quatro as variáveis que irão entrar inicialmente no teste: "Associação à administração em disputa", "Associação à administração em outra esfera", "postura acima da briga" e "uso do cargo". Consideramos que essas quatro características permitem a produção de um índice de governismo, onde zero é a ausência de governismo no HGPE do candidato e quanto maior o valor para o caso, mais governista foi o programa eleitoral do candidato. Rode o teste *α de Cronbach* no *RCommander*, interprete os coeficientes para o índice e interprete as alternativas de índice caso uma das variáveis fosse excluída do modelo.

3.5.2 A partir da tabela de distribuições a seguir, usando o pacote "CA" no *RStudio*, rode uma análise de correspondência canônica entre as variáveis "candidato e "índice de governismo". Também rode o gráfico de correspondência. Em seguida faça: a) análise dos coeficientes lembrando-se de que na análise de correspondência as duas variáveis são consideradas categóricas nominais. b) Interprete as distribuições das massas nas linhas e nas colunas. c) Interprete as capacidades explicativas das duas dimensões plotadas no gráfico. d) Faça uma tabela com o produto das massas para cada par de categorias e analise os resultados.

Tabela de distribuição do índice de governismo no HGPE do primeiro turno de 2014

Cand.	Sem Govern.	Baixo Govern.	Médio Govern.	Alto Govern.
Dilma	0	1	9	8
Marina	3	15	0	0
Aécio	0	16	1	1

3.5.3 Ainda utilizando do banco de dados BDCAP3V2_HGPE, inclua nele a variável índice de governismo que você criou na exercício 3.5.1, carregue o banco com a nova variável no *RCommander* e rode uma análise de correspondência múltipla com: índice de oposicionismo (já no banco de dados), índice de governismo (que você criou), sequência dia e proporção de tempo. O objetivo do teste é verificar se as áreas de algumas dessas variáveis correspondem no espaço. Explique os resultados.

3.5.4 Ainda utilizando o banco de dados BDCAP3V2_HGPE rode uma Análise de Componentes Principais usando as seguintes variáveis: Índice de oposicionismo, Índice de situacionismo, duração em segundos, ofensiva em relação a temas e sequência do dia. Essas cinco variáveis deverão fornecer os principais componentes da associação entre as categorias. Analise as distâncias e proximidades dos vetores, analise as contribuições para explicar as variações de cada dimensão e analise as contribuições individuais das variáveis para em cada dimensão.

3.5.5 Por fim, usando o banco de dados BDCAP3V2_HGPE rode uma análise de *clusters* para as variáveis: índice de oposicionismo, índice de situacionismo e ofensiva em relação a temas. No teste direto no *RCommander*, em Estatísticas e Análise Dimensional, rode primeiro pelo método *Kmeans* O objetivo é verificar como as variáveis de distribuem no espaço bidimensional. Em seguida, rode uma análise de *clusters* hierárquica pelo método Ward para produzir um dendrograma. Analise os resultados dos dois testes em conjunto.

ANEXO DO CAPÍTULO III

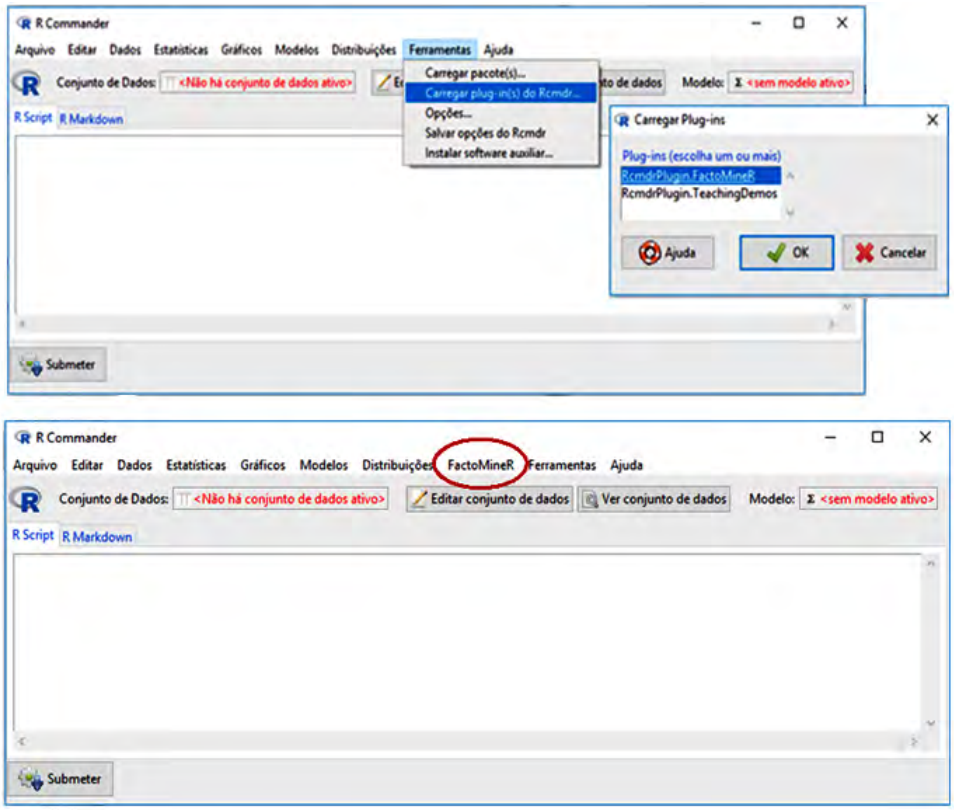
ANEXO 3.1 – INSTALAÇÃO DO PACOTE "CA" PARA ANÁLISE DE CORRESPONDÊNCIA NO *RSTUDIO*

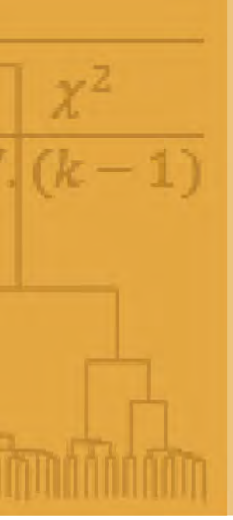
```
#INSTALAR O PACOTE "CA"

> install.package ("CA")


#CARREGAR O PACOTE "CA"

> library (CA)


#CARREGAR A TABELA DO EXCEL COMO BANCO DE DADOS A PARTIR DA
ÁREA DE TRANSFERÊNCIA ("CLIPBOARD"). PARA ISSO É PRECISO
SELECIONAR E COPIAR A TABELA DE DADOS.

> hgpe = read.table ("clipboard")


# LER A TABELA DE DADOS "HGPE" CARREGADO NA ÁREA DE TRABALHO
DO RSTUDIO

> hgpe

       não baixa media alta
dilma    0     1     9    8
marina   3    15     0    0
aecio    0    16     1    1


#SOLICITAR O TESTE DE ANÁLISE DE CORRESPONDÊNCIA PARA O BANCO
DE DADOS HGPE

> ca(hgpe)

Principal inertias (eigenvalues):
           1        2
Value      0.781291 0.078971
Percentage 90.82%   9.18%

 Rows:
           dilma    marina     aecio
Mass    0.333333  0.333333  0.333333
ChiDist 1.241126  0.819680  0.607057
Inertia 0.513465  0.223958  0.122840
Dim. 1  1.402999 -0.855432 -0.547567
Dim. 2  0.177746  1.126160 -1.303906
```

```
 Columns:
              não      baixa     media      alta
Mass     0.055556  0.592593 0.185185 0.166667
ChiDist  1.414214  0.641957 1.208305 1.186342
Inertia  0.111111  0.244213 0.270370 0.234568
Dim. 1  -0.967786 -0.713790 1.366595 1.342075
Dim. 2   4.007428 -0.421720 0.105264 0.046681


# SOLICITAR A GERAÇÃO DO GRÁFICO DE CORRESPONDÊNCIA PARA O
BANCO DE DADOS "HGPE"

> plot(ca(hgpe))
```

ANEXO 3.2 – INSTALAÇÃO DO *PLUG-IN* "*FACTOMINER*" NO *RCOMMANDER*

```
# INSTALAR O PACOTE NO RSTUDIO

>install.packages ("FactoMineR")


# CARREGAR O PACOTE INSTALADO NA ÁREA DE TRABALHO DO RSTUDIO

>library (FactoMiner)


# ABRIR O CONSOLE DO RCOMMANDER

>library (Rcmdr)


#NO CONSOLE DO RCOMMANDER ESCOLHER O CAMINHO
"FERRAMENTAS"/"CARREGAR PLUG-INS DO RCOMMANDER" PARA ABRIR A
CAIXA DE PLUG-INS INSALADOS NO RSTUDIO.

# SELECIONAR O PLUG-IN FactoMineR. O RCOMMANDER SERÁ
REINICIADO E NO MENU SUPERIOR APARECERÁ A OPÇÃO FactoMineR.
```





Capítulo IV

Análise de Conteúdo Aplicada a Redes Sociais Online

Análise de conteúdo é uma técnica de pesquisa que atribui categorias analíticas a conteúdos de comunicação, de acordo com regras prévias, usando métodos estatísticos (Riffe, 2014).

A Análise de Conteúdo é uma das mais antigas técnicas aplicadas a objetos empíricos no campo das pesquisas sobre os fenômenos sociais. Ela antecede o próprio campo de estudos da sociedade. Por se tratar de uma técnica aplicada à análise de textos, que se transforma ao longo do tempo, a Análise de Conteúdo demanda uma permanente atualização. Novos suportes físicos e tecnológicos para produção e difusão textual são, antes de tudo, um desafio aos pesquisadores que produzem ao mesmo tempo suas pesquisas e novos instrumentos de trabalho.

Existe uma profusão gigantesca de trabalho, manuais, textos explicativos e conceituais da Análise de Conteúdo aplicada a textos políticos. Riffe *et al.* (2004) fazem um levantamento exaustivo das pesquisas desenvolvidas nessa área. Não é objetivo de o capítulo fazer uma abordagem completa de toda a variação de instrumentos metodológicos ligados à Análise de Conteúdo (AC). Trata-se de uma proposta de AC automatizada aplicada a conteúdos publicados em Redes Sociais Online (RSO). A análise automatizada é uma parte da AC onde predomina o uso de programas informatizados e tecnologias que só se justificam para grandes conjuntos de textos. A automatização

mais atrapalha do que ajuda quando o número de textos a serem analisados é pequeno. As redes sociais online são um campo novo de produção e difusão de textos que nunca estão terminados. Além disso, os textos que circulam nas redes sociais oferecerem a possibilidade de uma medida estatística de impacto, chamado de engajamento, o que não era possível aferir a partir do próprio texto até as RSO.

Além de apresentar os fundamentos da Análise de Conteúdo por técnicas estatísticas, o capítulo também propõe uma técnica específica para análise de textos de RSO. Para tanto, em primeiro lugar, é apresentado o algoritmo proposto por Max Reinert como ferramenta analítica. A técnica é dividida em duas partes principais e dispensa o uso prévio de livros de códigos de termos e categorias léxicas para a classificação textual. Em segundo lugar, seguindo os demais capítulos deste manual, utiliza-se aqui para AC um *software* de código aberto como proposto no projeto GNU, que é uma interface do pacote estatístico R, chamada Iramuteq (Interface do R para Análises Multidimensionais de Textos e Questionários), desenvolvida pelo Laboratório de Estudos e Pesquisas Aplicadas às Ciências Sociais (LERASS)[1], da Universidade de Toulouse. Para acompanhar os exemplos e para fazer os exercícios propostos ao final do capítulo, você precisará da instalação da interface. O endereço para instalar do programa é: http://www.iramuteq.org/. Nele, também é possível acessar o passo a passo para instalação e os manuais oficiais do programa[2]. Também sugiro uma pesquisa em manuais e tutorais explicativos do uso do Iramuteq antes de entrar na parte empírica do capítulo.

4.1 Histórico da Análise de Conteúdo

O primeiro momento do que podemos chamar da fase científica da Análise de Conteúdo (AC) se dá nas primeiras décadas do século XX, nos Estados Unidos, quando pesquisadores – principalmente da Universidade de Columbia – passam a dar mais atenção ao rigor científico para interpretar textos de jornais. São definidas medi-

[1] Para acesso direto à página do grupo de pesquisa, acessar o link https://www.lerass.com/

[2] Por ser uma interface, o Iramuteq depende da instalação anterior do pacote estatístico R.

das comuns para medir o sensacionalismo nos artigos de periódicos, comparando os conteúdos de jornais de cidades grandes com os de pequenas localidades. Além da identificação de palavras-chave, passam a ser codificadas as características externas ao texto, como tamanho em número de caracteres ou espaço ocupado em cm2, localização na página, dia da semana em que o texto é publicado e o próprio jornal, que passaram a contextualizar a relevância do texto para os editores e, potencialmente, os efeitos para os leitores.

No início dos anos 1920, a atenção dos pesquisadores migra dos textos jornalísticos para as peças de propaganda. É nesse momento que as técnicas de Harold Lasswell ganham importância na análise dos conteúdos publicitários impressos nos Estados Unidos, em especial nos grandes centros urbanos, formados a partir do final do século XIX.

A segunda etapa dos estudos de AC se dá entre os anos 1940 e 1950, nos Estados Unidos, quando, junto com a manutenção dos estudos de conteúdos sociais e políticos, a técnica passa a ser usada em outros suportes textuais que não apenas jornais e peças publicitárias, tais como conteúdos transmitidos por emissoras de radiodifusão, discursos públicos de políticos, líderes religiosos e outros. Após a segunda guerra mundial, o governo norte-americano passa a usar a AC para identificar a existência de propaganda subversiva em jornais dos Estados Unidos. Para isso, são usadas listas de palavras consideradas chave na propaganda nazista. Aplicando a mesma sistematização da AC, White (1947) faz análise do romance autobiográfico de Richard Wright “Black Boy”, que trata da segregação racial nos Estados Unidos. O pesquisador usa três categorias para classificar as diferentes partes do texto, os fins e objetivos no texto, as normas e as pessoas. Com isso, ele é capaz de contextualizar os fatos relatados no livro com o contexto social norte-americano da primeira metade do século XX.

Bardin (1977) lembra que nas primeiras décadas do século XX os pesquisadores do campo da linguística e os cientistas sociais que analisavam conteúdo dos meios de comunicação se ignoravam. A não integração entre os dois campos teria gerado condições para o afastamento entre a Análise de Discurso, mais ligada ao campo linguístico, em busca das motivações de quem produz o texto em um *corpus* restrito, da Análise de Conteúdo, ligada ao campo das ciências sociais, cujo objetivo é descrever o texto produzido, normalmente a partir de um grande volume de material. Berelson

(1984) define a Análise de Conteúdo como uma técnica de pesquisa que tem por finalidade a descrição objetiva, sistemática e quantitativa do conteúdo manifesto em qualquer suporte de comunicação. A partir desta definição é possível diferenciar, sem sombra de dúvidas, os objetivos e a forma de pesquisa da Análise de Conteúdo e da Análise do Discurso. A técnica apresentada neste capítulo se aplica à Análise de Conteúdo e não a de Discurso.

Se até os anos 1950 temos a consolidação das técnicas de AC com replicação delas em diferentes suportes textuais, na década seguinte, quando começa a terceira etapa de desenvolvimento desse tipo de pesquisa, há uma diversificação de abordagens e técnicas para estudo de conteúdos. Berelson (1984), reconhecendo as limitações da AC, afirma que o método não possui qualidades mágicas e não é possível extrair do texto mais do que ele apresenta explicitamente, ainda que a Análise do Discurso defenda o contrário. Há uma reconsideração sobre o uso da Análise de Conteúdo para fins de controle e antecipação de demandas sociais por parte de governos. Com o "retorno" das técnicas de AC para finalidades acadêmicas e não mais políticas e comerciais, três pontos fundamentais são desenvolvidos nas décadas seguintes. O primeiro é a entrada do processamento de dados por computadores pessoais, com programas específicos para a AC, o que aumentou o rigor analítico e a amplitude das técnicas utilizadas. O segundo é uma passagem da simples descrição de ocorrências em um texto para a busca da inferência analítica, em parte, retomando os objetivos do uso da AC para antecipar interpretações e efeitos no público. E, por fim, uma ampliação do uso da técnica para outros campos de conhecimento, de forma compartilhada, como não acontecia antes (Bardin, 1977). Um exemplo é a utilização de AC em processos judiciais, em especial na área criminal, como prova de autoria a partir de textos com autoria identificada e textos anônimos.

A transição do século XX para o XXI apresenta novos desafios e grandes potencialidades para o desenvolvimento da AC, com a aplicação da técnica à análise de conteúdos no suporte digital – internet em geral e mais especificamente as Redes Sociais Online (RSO). Além de permitir acesso a um conjunto de textos produzidos especificamente ou apenas disponibilizados na internet, os meios digitais ampliaram exponencialmente as possibilidades analíticas com a difusão das RSO. É que nesse tipo de rede

social, além do texto original produzido, há também interação e (re)produção de novos textos por parte do público que acessa os conteúdos iniciais.

Ao invés de novos problemas, as tecnologias digitais são desafio para a AC, que desde sua origem aplica-se a qualquer tipo de comunicação humana. Para enfrentar esse desafio é que a Análise de Conteúdo reforça seu objetivo inicial: permitir que se ultrapassem incertezas sobre o que está efetivamente contido no documento, ficando restrita à categorização de termos, léxica, ou, no máximo, em sentidos extraídos do próprio texto e a enriquecer a leitura dos conteúdos, evitando uma descrição simplista, imediata e espontânea do próprio texto. O grande volume, a temporalidade e a fragmentação dos conteúdos presentes nas redes sociais online reforça a posição de que a AC continua sendo o método empírico mais adequado para análise e interpretação de manifestações humanas registradas em algum suporte comunicacional de massa.

Uma definição atualizada da AC, apresentada no último quarto do século passado, segundo Bardin (1977), é a de que se trata de um conjunto de técnicas de análise de comunicações que tem como objetivo a descrição sistemática de conteúdos das mensagens a partir de indicadores quantitativos que permitem fazer inferências sobre as condições de produção e, por vezes, da recepção dessas mensagens. Os conteúdos de RSO são propícios para integrar as descrições de textos originais, ou seja, aqueles que deram origem às interações, aos textos que resultaram das conversações e interações a partir do conteúdo inicial, o que permite uma nova forma de produção de inferências aos conteúdos. Além disso, as RSO também produzem uma medição objetiva do impacto geral de determinado conteúdo em função das interações positivas ou negativas, compartilhamentos, etc.

Embora seja uma técnica centenária de pesquisa, a Análise de Conteúdo tem acompanhado as transformações tecnológicas e, no estágio atual, um ramo que mais se desenvolve é o da Análise de Conteúdo Automatizada, aquela que está diretamente ligada às capacidades de captura, organização, interpretação e análise apoiadas por dispositivos tecnológicos e por pacotes estatísticos (Grimmer & Stewart, 2013). A proposta de análise empírica apresentada aqui se filia ao conjunto de técnicas de Análise de Conteúdo Automatizada aplicada às interações textuais em Redes Sociais Online (RSO), como veremos nos próximos tópicos.

4.2 Etapas da Análise de Conteúdo aplicada a textos políticos

A Análise de Conteúdo é uma técnica específica voltada para identificação, descrição e predição de elementos textuais. Ela é predominantemente descritiva, portanto, frequentista por natureza. Trata-se de uma epistemologia realista, de acordo com Lombard *et al.* (2002); Neuendorf (2016); e Drisko (2016), onde a interferência ou subjetividade do pesquisador deve ser a menor possível. O ponto forte da técnica de análise de conteúdo está na confiabilidade e na validade dos resultados obtidos a partir da interpretação das características do texto. Por isso, há uma forte dependência entre a relação do que está no texto e o que isso significa para quem está realizando a pesquisa.

Krippendorff (2004) lembra que a análise de conteúdo sempre se refere a algum tipo de manifestação humana passível de ser transmitida e compreendida. Os textos são produzidos por alguém pensando nas possíveis interpretações dos seus leitores, mas eles não possuem qualidades objetivas próprias, ou seja, eles não têm significados em si mesmos. O significado depende de quem interpreta, ou seja, quem tem acesso a eles. Por este motivo que um texto nunca tem um significado único (Krippendorff, 2004; Riffe *et al.*., 2014). Cabe ao pesquisador encontrar, identificar e descrever os objetos que podem dar origem a diferentes interpretações, dependendo da perspectiva de quem interpreta os conteúdos. Em outras palavras, o significado de um texto nunca está presente nele mesmo e sim nos seus leitores. Mais do que isso, os significados específicos podem mudar, dependendo do contexto em que determinado conteúdo é produzido e consumido (Krippendorff, 2014).

Mesmo com a utilização de técnicas objetivas de captura e tratamento de informações de textos, o pesquisador precisa controlar os efeitos da interpretação pessoal dos significados dados aos elementos textuais. Historicamente, a análise de conteúdo é feita a partir da leitura sistemática para extração de partes representativas dos textos, permitindo a redução dos dados para representar estruturas de conteúdos (Alonso, 2012). Ela pode ser agrupada em três grandes dimensões analíticas:

1) Interna ao próprio texto – interessada na forma como se compõem e quais são os elementos predominantes do texto;

2) Das causas para produção de um texto com determinadas características –

interessada no contexto que envolve a produção do texto, ou seja, no entorno e não no texto em si, e;

3) Análise dos efeitos de dados conteúdos – interessada no contexto que envolve a apropriação e interpretação do texto, ou seja, na recepção dos conteúdos.

A forma mais comum de AC é a análise interna, aquela pesquisa cujo objetivo é descrever e fazer inferências a partir do que está explícito no texto, após sua codificação e organização em categorias analíticas. A partir dela, e tendo conhecimento das características contextuais para a produção e de quem o produziu, é possível extrapolar do texto em si para as causas à sua produção.

A unidade de codificação mais comum em análise de conteúdo é a palavra ou termo. Mas também pode ser a frase ou outras unidades menos comuns. Quando o objetivo principal é analisar a presença e intensidade de léxicos, a análise léxica, o objeto principal é a palavra ou o termo. Se o objetivo for interpretar sentidos, via análise semântica, o objeto preferencial de análise é a frase ou qualquer unidade de texto que permita extrair algum sentido de cada trecho do *corpus* textual. O processo de codificação é fundamental para garantir a máxima consistência nas representações, alcançando as mínimas validade e confiabilidade na pesquisa (Alonso, 2012).

Com o objetivo de sistematizar o uso recente da técnica de análise de conteúdo nas ciências sociais, Bardin (1977) divide em três etapas o avanço da técnica no século XX[3]. Em todas elas as perguntas centrais para a análise de conteúdo estão presentes: o que posso analisar? O que é passível de ser interpretado? Para responder as questões, é preciso analisar as unidades básicas dos conteúdos dos textos. Assim, análise de conteúdo não é doutrinal, nem normativa. Ela não visa encontrar o que pode estar encoberto a partir do texto, de forma intuitiva, mas sim o que o texto apresenta explicitamente.

[3] Entre as primeiras experiências de Análise de Conteúdo que se tem notícia, Bardin (1977) cita a análise de 90 hinos religiosos na região em que hoje se encontra a Suécia para identificar se eles teriam um efeito negativo sobre os luteranos. Foi a primeira vez em que se tem notícia do uso das categorias Favorável/Desfavorável para trechos de textos. Em 1888, o francês B. Boudon faz análise das emoções presentes no conteúdo do livro bíblico Êxodo a partir da indicação de presença de palavras-chave no texto. Já no início do século XX, em 1908, uma equipe formada por pesquisadores da Universidade de Chicago e por professores poloneses sistematiza a leitura de cartas, diários e artigos de jornais sobre imigrantes poloneses nos Estados Unidos e no país de origem (Bardin, 1977).

4.3 Descrição da proposta de análise em duas etapas com método Reinert

O objetivo específico do capítulo é apresentar uma proposta de Análise de Conteúdo Automatizada com menor interferência possível de subjetividades do pesquisador na categorização textual. Na técnica apresentada aqui, a unidade de análise é o termo/palavra isolado. Mede-se a presença total (número de citações do termo/palavra), a presença relativa por sub*corpus* do texto (citações do termo/palavra no *cluster* temático) e as relações com outros termos/palavras (presença em diferentes *clusters*).

Um objetivo que deriva do anterior é criar um *corpus* textual que apresente determinada caraterística (por exemplo, trata de um ou mais assuntos públicos) do total de textos, utilizando para isso um método que independa da subjetividade do pesquisador ou de categorizações prévias ao texto. Para isso, a primeira parte da qualificação proposta é pelo método de Reinert (1990), conhecido por Classificação Hierárquica Descendente (CHD). A segunda é a classificação das partes do texto que contém os termos identificados como estatisticamente significativos, também segundo Reinert (1990). Trata-se de um método que apresenta uma classificação hierárquica descendente das ocorrências dos termos em um segmento específico do texto. Assim, estamos fazendo duas coisas ao mesmo tempo: a primeira é identificar coocorrências de termos nos mesmos segmentos, distribuindo textos em classes por proximidade; a segunda é hierarquizar a presença relativa de cada termo nas classes de palavras. Como exemplo aqui, a técnica é aplicada a um conjunto de textos políticos publicados por pré-candidatos em suas páginas oficiais em uma RSO no período do mês de maio de 2018.

A classificação hierárquica descendente é uma das técnicas mais importantes para a análise léxica automatizada de conteúdos de textos e documentos. Ela parte da lógica da existência de correlação entre termos dentro de um mesmo segmento de *corpus* textual. A definição dos limites do *corpus* textual e a mediação da intensidade de presença dos termos em diferentes *corpus* permite identificar possíveis associações entre termos por proximidade e intensidade. Max Reinert criou o algoritmo usado incialmente pelo *software* Alceste. Depois, o mesmo algoritmo foi introduzido na interface com código aberto para o pacote estatístico “R”, o Iramuteq. A fundamentação teórica do algoritmo de Reinert é inspirada nas propostas de Benzécri (1992), para análise léxica. Ela consiste em testar

leis de distribuição de vocábulos em um *corpus* textual qualquer. Não se trata de uma análise sintática, mas da verificação dos termos presentes nos textos, da forma como eles se organizam e dos elementos constitutivos deles. Isso é o que Reinert (1990) chama de análise dos "mundos lexicais".A proposta de Reinert (1990) permite um avanço nas descrições, passando da simples presença e quantidade de léxicos para uma associação com o contexto da presença de termos. Em função de limitações para o uso do nome Alceste, que é um *software* com código fechado, a interface do pacote estatístico "R" para análise textual, Iramuteq, incorporou o algoritmo de análise por CHD, porém, dando-lhe o nome do autor, Reinert. Trata-se do mesmo algoritmo para classificação hierárquica, porém, com o nome do precursor do método de análise (Camargo, 2013).

Normalmente, desenhos de pesquisa que usam a técnica de análise de conteúdo partem de um conjunto de categorias ou variáveis pré-estabelecidas, na maioria das vezes, organizadas em um livro de códigos que visa guiar a ação do pesquisador na busca de termos que representem as características estudadas. O problema é que, sendo os livros de códigos preparados antes da coleta de dados – e muitas vezes para diferentes conjuntos de textos –, não há garantia de que eles serão capazes de capturar todas as especificidades do texto em análise. Por isso, aqui a proposta de análise inverte o processo. Busca-se de saída, no próprio *corpus* empírico, através do uso do algoritmo de Reinert, a identificação dos termos que mais aparecem e que se aproximam entre si nos textos e, portanto, formam classes de termos com homogeneidade interna. Para isso, a técnica parte dos seguintes pressupostos:

1 – A identificação de classes de palavras sem prévia codificação de categorias tem dupla finalidade:

1.a) Separar o conjunto de textos sobre temas dos candidatos e das campanhas. PRESSUPOSTO: os textos são distintos. No exemplo apresentado aqui, as publicações que tratam de assuntos da gestão, transporte, saúde, etc. não fazem ataques ou defesas das imagens de gestões.

1.b) A partir do *cluster* de termos que tratam de temas, agregar palavras com significado exclusivo de tema em conjuntos de temáticas mais amplas. PRESSUPOSTO: o conjunto de termos aumenta o nível de significados do debate em relação aos termos isolados.

2 – No banco de dados com os textos, criar variáveis que indiquem presença ou ausência dos termos/conjuntos nas manifestações, para, posteriormente, realizar testes estatísticos que permitam identificar associações entre as categorias e, no caso de RSO, associações entre as categorias e o impacto que os textos tiveram em termos de interações.

Com isso, tem-se como principal resultado a produção de categorias analíticas que não são *ad-hoc*, mas o produto de um duplo processo de filtragem: i) pelo *cluster* CHD, que distingue o conjunto de termos temáticos dos demais conjuntos; e ii) pela frequência relativa da presença do termo no *cluster* temático. A seguir, são apresentadas as etapas da técnica de análise textual proposta aqui a partir da análise por CHD de Reinert.

4.4 O método Reinert na análise de conteúdo de Redes Sociais Online

Para a aplicação do algoritmo de Reinert à identificação de *clusters* e níveis de interação, a técnica proposta aqui neste capítulo passa por três etapas:

1º - Rodar o teste de *cluster* usando o algoritmo de Reinert.

2º - Identificar no *cluster* temático quais são os termos mais frequentes.

3º - No banco de dados com os textos criar variáveis que agreguem termos que representem o mesmo tema ou áreas afins.

Além de proporcionar uma técnica automatizada de análise de conteúdo, há um processo interno de validação a partir da seleção e textos que contenham as temáticas que geraram os *clusters*. Se forem criados novos agrupamentos a partir destes subconjuntos de textos, a validação indica ser necessária outra filtragem, para diferenciação de conteúdos. Se não houver separação dos termos em *clusters* é porque os textos selecionados, embora a partir de temas distintos para o *cluster*, fazem parte de um sub*corpus* textual que apresenta unidade interna.

Apesar da validação automatizada, o método apresenta algumas limitações para as análises. A primeira delas é que as categorias precisam ser mutuamente excludentes (ponto de partida para criação de *clusters*). Por exemplo, no caso do estabelecimento de uma categoria de análise como “política pública”, os temas específicos de política pública não são mutuamente excludentes, pois é possível que exista um texto

com política pública de educação e de segurança ao mesmo tempo. Aqui, educação e segurança passam a ser subcategorias de uma classe textual mais ampla – que serão diferenciadas na terceira etapa descrita acima. A segunda limitação para a análise automatizada é que se o conjunto de textos não permite a diferenciação em classes (os *clusters*), a primeira etapa do processo torna-se pouco útil. Ou seja, o texto não permite uma categorização temática a partir de suas próprias características – nesses casos é necessário usar uma classificação prévia, estabelecida pelo pesquisador.

Como toda técnica aplicada, a proposta de análise automatizada em duas etapas apresenta limitações, como qualquer ferramenta metodológica de pesquisa. A principal limitação é a impossibilidade de replicar exatamente as mesmas categorias em diferentes conjuntos de textos. Afinal, a categorização será definida em função dos termos presentes em cada conjunto de texto e não previamente. Outro ponto a se considerar na aplicação do método automatizado é a heterogeneidade interna do texto. Faz mais sentido fazer análise de conteúdo em duas etapas em textos coletivos e com liberdade de forma e temática do que em textos mais formais e homogêneos em termos temáticos. No último caso, é possível conhecer previamente os temas que aparecerão na Análise de Conteúdo, portanto, sendo dispensável a abordagem exploratória inicial. Por estes dois motivos é que a proposta se aplica principalmente aos textos publicados em grupos ou páginas de redes sociais online. Apesar das limitações, apresentaremos a seguir as vantagens comparativas da AC automatizada.

Para demonstrar a proposta de análise de conteúdo automatizada aqui, será usado o conjunto de postagens feitas em *fanpages* no Facebook de quatro figuras públicas brasileiras que em maio de 2018 eram pré-candidatos à Presidência da República. São as *fanpages* de Geraldo Alckmin (PSDB), Ciro Gomes (PDT), Jair Bolsonaro (PSL) e Lula (PT). Foram coletadas todas as postagens publicadas nas quatro páginas oficiais destes políticos entre 1º e 31 de maio de 2018. Ao todo, são 374 publicações no período. Sendo 75 de Alckmin, 54 de Bolsonaro, 59 de Ciro Gomes e 185 de Lula. Como se percebe, Lula teve uma atividade mais intensa que os demais candidatos no Facebook no período. O que explica isso é que o mês de maio é o primeiro que o ex-presidente passou na prisão por ação criminal proveniente da Operação Lava Jato.

Pelo método tradicional de organização do *corpus* empírico para Análise de

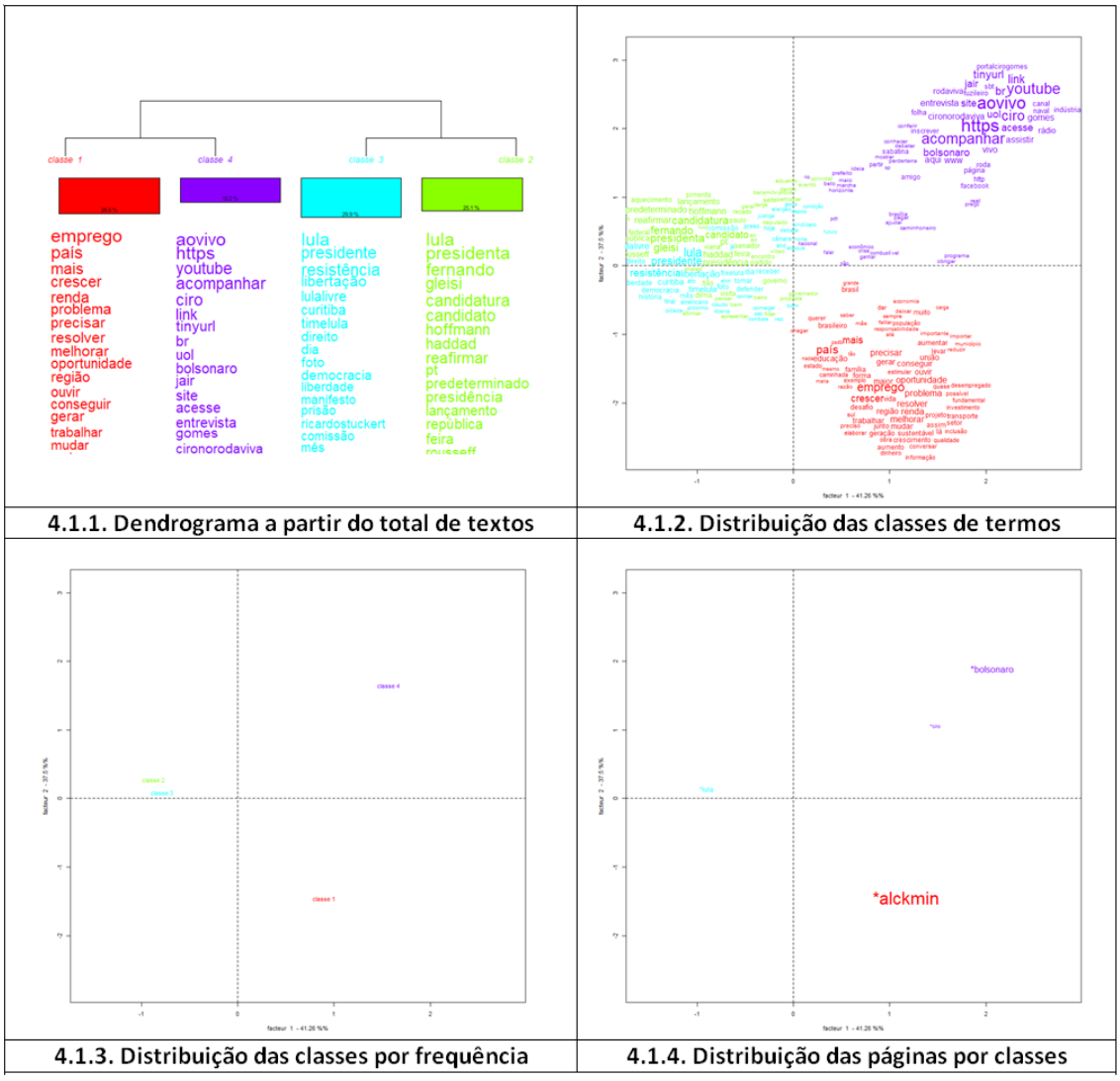
Conteúdo, o pesquisador teria uma lista (livro de códigos) com os termos que representam determinados assuntos e anotaria a presença ou ausência deles nas postagens. Em seguida, agregaria, verificaria a proporção de presença e a distribuição dela nos textos postados nas páginas de cada um dos pré-candidatos. Isso implicaria que cada página seria uma classe de análise, pois se sabe que as origens dos conteúdos são diferentes por página. No método proposto aqui, não há um livro de códigos prévio. Todos os textos são organizados em um único *corpus* e, após a aplicação do algoritmo de Reinert, os subconjuntos de termos são reagrupados por intensidade e proximidade de citações. Esta é a primeira etapa da análise. Em seguida, conhecendo as classes de textos, o pesquisador deve voltar ao *corpus* para classificá-lo e acrescentar novas variáveis descritivas das características.

Assim, o primeiro passo da análise automatizada é identificar os *clusters* para o conjunto total de textos. O gráfico 4.1 a seguir mostra a formação de quatro categorias no conjunto de 374 postagens dos pré-candidatos. O dendrograma do lado direito (imagem 4.1.1) mostra a presença hierárquica dos termos nos quatro *clusters*, além de indicar que as classes 1 e 4 estão mais próximas entre si, enquanto as classes 2 e 3 fazem parte de outro grupo maior. Além do número de classes, também é possível perceber a participação percentual delas no total de textos. A maior é classe 3, com 29,9% dos termos válidos, seguida da classe 1, com 26,8%; classe 4, 15,1%; e classe 2, com 18,2%. Como se percebe, os tamanhos das classes são relativamente homogêneos, com pouco mais da metade dos termos em cada um dos dois grandes conjuntos (classes 1 e 4 de um lado, classes 2 e 3 de outro). Visualmente, também é possível perceber que os *clusters* não reproduzem as *fanpages*. A classe 1 apresenta palavras relativas ao tema "Emprego/economia". A classe 2 mostra termos ligados a espaços de comunicação externos ao Facebook, tais como *youtube, link, entrevista*, além dos nomes de *Bolsonaro* e *Ciro*. As outras duas classes começam com o termo *Lula*, indicando que se relacionam ao nome do ex-presidente, porém, com estruturas hierárquicas de termos distintas.

A distribuição dos termos no espaço bidimensional (imagem 4.1.2) mostra como os *clusters* se organizam internamente e as distâncias entre eles, além de indicar a intensidade da presença dos termos em cada categoria a partir do tamanho das palavras. Nos quadrantes da parte esquerda, o gráfico 4.1.2, indica que as categorias 2 e 3

não apresentam diferenças espaciais, ou seja, estão nos mesmos quadrantes e muito próximas entre si (termos com cores sobrepostas). Na parte superior direita está a classe 4, com os termos relativos a dois pré-candidatos e a atividades como *acompanhar, entrevistas, ao vivo*, etc. Na parte inferior direita aparece a classe 1, com as palavras relativas ao grupo temático das postagens.

Gráfico 4.1. Distribuição de termos por classes pelo método Reinert



4.1.1. Dendrograma a partir do total de textos | 4.1.2. Distribuição das classes de termos

4.1.3. Distribuição das classes por frequência | 4.1.4. Distribuição das páginas por classes

Fonte: autor.

O gráfico 4.1.3 acima mostra as posições, proximidades e distâncias dos quadrantes para as classes CHD. A classe 1 fica no quadrante inferior direito e a classe 4 no superior direito. Isso indica a proximidade maior entre as duas do que com as demais. Já as classes 2 e 3 localizam-se muito próximas entre si, no mesmo quadrante. O quarto

gráfico (4.1.4) aponta as distribuições das páginas no espaço, segundo os termos encontrados nas classes. Percebe-se que a página de Alckmin coincide com a classe 1, a página de Lula está no espaço das classes 2 e 3, enquanto as páginas de Bolsonaro e Ciro Gomes compartilham o espaço da classe 4.

Feitas as descrições dos resultados do CHD, o segundo passo da análise é identificar pelas frequências relativas os termos que mais aparecem em cada *cluster* temático e agrupá-los em conjuntos que façam sentido analítico. O dendrograma na imagem 4.1.1 distribui os termos por critério de proximidade. O método Reinert, além de indicar as classes, também permite identificar que palavras aparecem mais em um *cluster* do que nos demais, ou seja, quais os termos que são específicos de cada um dos grupos. Isso é feito a partir da estatística χ2 de Pearson, que indica se a presença de um termo em um *cluster* é estatisticamente diferente da presença do mesmo termo em outros *clusters*. Assim, podemos usar a distribuição hierárquica de termos que apresentam significância estatística para identificar cada uma das quatro classes aqui, conforme o quadro 4.1 a seguir. A linha de % de ocorrências mostra a participação relativa de cada classe no total de textos classificados pelo algoritmo. A classe 3 (Lula réu) é a maior, com 29,8% do total de ocorrências. Depois vem a classe 1 (Emprego), com 26,8%, seguido de perto da classe 2 (Lula candidato), com 25,1%, e, por fim, a classe 4 (campanha), com 18,8% do total de ocorrências dos textos. A linha "nome das categorias" serve para que o pesquisador apresente um termo que represente o significado do conjunto de termos de cada classe.

Quadro 4.1. *Clusters* de palavras para criação de categorias de análise

	Classes produzidas pelo algoritmo de Reinert			
	CLASSE 1	**CLASSE 2**	**CLASSE 3**	**CLASSE 4**
Termos com χ^2 significativo por classe	Emprego, país, crescer, renda, problema, resolver, oportunidade, gerar, trabalhar, mudar, união.	Lula, presidente, Fernando haddad, gleisi Hoffman, candidato, reafirmar, PT, presidência, lançamento, república.	Lula, presidente, resistência, libertação, lulalivre, Curitiba, timelula.	Ao vivo, youtube, acompanhar, ciro, link, uol, Bolsonaro, site, acesse, entrevista, assistir, roda viva, rádio.
% de Ocorrências	26,8%	25,1%	29,8%	18,8%
Nomes das categorias	Emprego	Lula candidato	Lula réu	Campanha
% de Postagens	17,5%	35,4%	33,7%	13,5%
% de casos	30%	60,7%	57,8%	23,1%

A linha do % de postagens mostra a distribuição das categorias no conjunto de 374 textos. São 35,4% para Lula Candidato, 33,7% para Lula réu, outros 17,5% para o tema Emprego, único tema de política pública entre as categorias geradas pelo CHD, e 13,5% para o tema Campanha. Como um *post* pode estar presente em mais de uma categoria, a linha % de casos mostra o total de presença por categoria. As duas envolvendo o ex-presidente Lula são as que mais aparecem, seguidas de emprego e depois campanha. Com isso é possível identificar que temas tiveram mais presentes no conjunto de textos que fazem parte da análise. O próximo passo da técnica é classificar os textos em função das posições deles nas classes. Porém, antes disso, para fins didáticos, vamos comparar no próximo tópico qual seria o resultado da aplicação da AC aos mesmos textos de *fanpages*, porém usando um livro de códigos definido previamente. O livro de códigos vem sendo aplicado nas últimas eleições a textos políticos produzidos por meios de comunicação, em jornais, websites noticiosos e weblogs, e aos textos produzidos pela elite política durante as campanhas, especialmente no Horário Gratuito de Propaganda Eleitoral (HGPE).

4.5 Uma comparação com o método tradicional de classificar textos políticos

Para mostrar as diferenças dos resultados entre os métodos de classificação, vamos classificar as mesmas postagens a partir dos léxicos presentes em um livro de códigos usado em Análises de Conteúdo (AC) de textos eleitorais há mais de uma década pelo grupo de pesquisa em Comunicação Política e Opinião Pública (CPOP)[4]. A utilização de livros de códigos prévios ao texto é tradicionalmente usada em análises de conteúdo há um século. Segundo essa técnica, a partir de um conhecimento empírico prévio ou extraído da literatura especializada no tema, categorizam-se determinados atributos do texto. Aqui, a característica do texto a ser mensurada pela AC é o seu tema. Por exemplo, é montada uma lista de léxicos que representem temas de

[4] Para ver o conjunto de trabalhos produzidos ao longo do tempo pelo CPOP, acessar: http://www.cpop.ufpr.br/sobre-o-cpop/

política pública. Com base nesta lista e após a leitura dos textos, é feita a classificação nas diferentes categorias. No caso das publicações em Facebook, a presença de um termo classificaria a postagem como sendo relativa a um tema. A presença de dois ou mais termos permite o posicionamento do *post* na categoria "cardápio", quando trata de mais de um tema de política pública. Se a postagem não apresenta nenhum dos termos, ela não é classificada como válida.

A tabela 4.1 a seguir mostra a distribuição das presenças dos termos nos 374 *posts* das quatro *fanpages* analisadas aqui. A primeira informação que chama atenção é que menos de 25% das postagens apresenta pelo menos um dos termos presente no livro de códigos. Em 85,5% das publicações, não há presença de nenhum dos léxicos que representam os temas presentes no livro de códigos. Além disso, oito dos 18 temas não tiveram nenhuma citação no período em análise. O tema que mais aparece no conjunto de postagens é "Economia/emprego", com 23 *posts*, que representa 42,59% dos casos válidos e 6,16% do total. Em segundo lugar, empatados, ficam os temas "Educação" e "Cardápio", com 14,8% dos casos válidos e 2,14% do total.

A coluna de "termos/tema" equivale à lista prévia de palavras selecionadas para compor o livro de códigos. As três primeiras colunas da parte direita da tabela 4.1 mostram a distribuição por número de postagens e percentuais para as *fanpages* de candidatos. Na linha "sem termos", estão as postagens que não foram classificadas em nenhuma categoria temática presente no livro de códigos. Alckmin tem aproximadamente metade não classificada (56%). Os demais ficam acima de 90% de não classificação, o que indica pouco rendimento na classificação pelas categorias previamente estabelecidas.

Tabela 4.1. Distribuição dos termos a partir do livro de códigos do CPOP

TERMO/TEMA	N	%Válido	%Total	Alckmin	Bols.	Ciro	Lula
Economia/Emprego	23	42,59	6,16	14 (18,6)	1(1,7)	3(5,0)	5(2,7)
Educação	8	14,81	2,14	4(5,3)	1(1,7)	0	3(1,6)
Cardápio	8	14,81	2,14	6(8,0)	0	0	2(1,1)
Bem estar (bolsas)	4	7,4	1,07	1(1,3)	0	0	3(1,6)
Saúde	2	3,7	0,53	2(2,6)	0	0	0
Tributária	2	3,7	0,53	2(2,6)	0	0	0
Criança	2	3,7	0,53	1(1,3)	0	1(1,7)	0
Corrupção	2	3,7	0,53	0	0	1(1,7)	1(0,5)
Segurança Pública	1	1,85	0,26	1(1,3)	0	0	0
Transporte	1	1,85	0,26	1(1,3)	0	0	0
Funcionalismo	1	1,85	0,26	1(1,3)	0	0	0
Infraestrutura e saneam. básico	0	0	0				
Desenv. Urbano, planej. Urbano	0	0	0				
Esporte/cultura/lazer	0	0	0				
Meio Ambiente	0	0	0				
Orçamento	0	0	0				
Idoso	0	0	0				
Mulher	0	0	0				
Agricultura	0	0	0				
Sem Termos	319	100	85,52	42(56)	52(96,6)	54(91,6)	171(92,5)
Total	373		100				

Fonte: autor

Analisando as colunas que indicam as distribuições dos candidatos, a diferença é que a *fanpage* de Alckmin é a que apresenta o maior percentual de postagens temáticas. Os outros três políticos apresentam *posts* que não tratam de temas de política pública, ainda que publiquem tanto quanto Alckmin no período analisado. Dado o baixo rendimento do livro de códigos inicialmente proposto e a impossibilidade de aceitar um resultado tão baixo de postagens classificadas, seria necessário revisar o livro de códigos para incluir termos que não foram elencados inicialmente. Ao fazer isso, os pesquisadores descobririam que as postagens não são majoritariamente temáticas porque no período analisado os políticos dedicaram suas *fanpages* a outros conteúdos, que não de políticas públicas, com a exceção de "Emprego" – conforme descrito no quadro 4.1 acima. A partir do próximo tópico, retomaremos a técnica automatizada de AC. Ou seja, o método proposto aqui, além de utilizar o próprio *corpus* empírico para definir as categorias analíticas, evita que o trabalho de coleta de dados tenha que ser refeito por imprecisão de termos presentes no livro de códigos.

A justificativa para abrir mão da categorização prévia encontra-se na necessi-

dade de ajustar o máximo possível as categorias analíticas ao conteúdo do texto. Isso é fundamental quando se considera o conjunto de manifestações em Redes Sociais Online, que são mais heterogêneas e abertas, com menor controle prévio do que será publicado. A comparação feita aqui (ver tabela 4.1), mostra que usando uma categorização prévia, sem ajuste, o percentual de postagens sem classificação seria muito alta, ultrapassando 90% dos textos em três das quatro páginas analisadas. Além disso, metade das categorias previstas no livro de códigos não tem uma única ocorrência no conjunto de textos analisados. Evidente que se torna necessária a adaptação das categorias para aumentar o rendimento das categorias analíticas. No entanto, isso ainda seria insuficiente, pois como a técnica automatizada indicou, das quatro categorias produzidas, apenas uma é temática (economia). Outras três indicam a utilização de textos para debates de assuntos que não são temas públicos e dificilmente seriam identificados em categorias analíticas tradicionais da Análise de Conteúdo de textos políticos.

Para aproximar as categorias de conteúdo dos textos analisados, a proposta é usar um algoritmo para determinar quais termos estão mais presentes e mais próximos ou distantes entre si. Trata-se do algoritmo proposto por Reinert, usado na primeira etapa, para identificar o número de *clusters* e que termos compõem cada classe, para posterior análise. No caso utilizado aqui, o conjunto de postagens de quatro páginas de pré-candidatos a presidente gerou quatro classes de termos com afinidade entre eles. A partir disso, as classes foram assumidas como categorias temáticas na variável “tema geral”, criada posteriormente aos resultados obtidos pelo método Reinert. A categorização dos textos mostrou a existência de apenas um *cluster* temático, formado por léxicos que remetem a temas da economia. Os outros três conjuntos não são temáticos. Dois deles envolvem Lula, um com léxicos sobre a pré-candidatura e política, outro sobre as acusações criminais contra o ex-presidente. Por fim, a quarta categoria remete a um tema de divulgação de campanhas fora da rede social, com chamadas e links para entrevistas e participações em eventos públicos. O principal achado a partir do método automatizado é que a maior parte dos temas diz respeito a outros assuntos, que não temas de política pública.

4.6 Análises das classificações a partir da tematização automatizada

Voltando ao método automatizado em duas etapas proposto aqui, o próximo passo é criar as categorias temáticas no banco de dados a partir dos termos identificados hierarquicamente nos *clusters*, conforme a tabela 4.1 acima. Assim, voltamos ao banco de dados com os textos das postagens para criar categorias analíticas a partir do conjunto de termos de cada *cluster*. A imagem 4.1 a seguir reproduz parte do banco de dados usado aqui. Os textos foram capturados das páginas dos políticos usando a ferramenta "Netvizz", proposta por Rieder (2013), para extrair dados da RSO principalmente fins acadêmicos utilizando a API disponibilizada pelo Facebook. Perceba que na variável "post_mesage" constam os textos originais. As colunas de CL1 a CL4 indicam se as palavras que aparecem em cada *cluster* com significância estatística constam na postagem. Perceba que existem postagens sem se classificar em nenhuma, em uma, duas ou três classes, como indicado na coluna "soma".

Imagem 4.1 Segmento do banco de dados com os textos codificados

page	type	CLASTRAD	post_message	link_domain	post_published	post_published_sql	CL1	CL2	CL3	CL4	SOMA	likes_count_fb
Ciro	video	0	Voltamos com Ciro Gomes falando aovivo no I Congresso Nacional do PDT Div	facebook.com	2018-05-05T14:21:02+0000	05/05/2018 16:21	0	0	0	1	1	842
Lula	link	6	Você lembra. Com Lula não faltava emprego e você vivia melhor. timelula	noticias.r7.com	2018-05-29T19:07:14+0000	29/05/2018 21:07	1	1	1	0	3	2402
Bolson	photo	0	Vamos revogar isso em 2019. Faremos o Brasil grande.	facebook.com	2018-05-13T00:38:49+0000	13/05/2018 02:38	0	0	0	0	0	19726
Ciro	video	0	Vamos debater energia alternativa! Acompanhe o Ciro aovivo no evento sobre	facebook.com	2018-05-30T17:14:24+0000	30/05/2018 19:14	0	0	0	1	1	706
Ciro	video	0	Vamos debater energia alternativa! Acompanhe o Ciro aovivo no evento sobre	facebook.com	2018-05-30T17:41:51+0000	30/05/2018 19:41	0	0	0	1	1	872
Lula	photo	0	Uma multidão de Lulas tomou as ruas de Buenos Aires capital da Argentina n	facebook.com	2018-05-20T00:03:08+0000	20/05/2018 02:03	0	1	1	0	2	11287
Alckmi	photo	0	Uma grande honra participar há pouco da inauguração da sede do PSDB/MA.	facebook.com	2018-05-05T12:54:36+0000	05/05/2018 14:54	0	0	1	0	1	637
Alckmi	status	3	Uma das maiores injustiças do Brasil é a falta de educação de qualidade para todos. Não che	facebook.com	2018-05-24T15:45:47+0000	24/05/2018 17:45	1	0	0	0	1	1109
Alckmi	photo	17	Uma das coisas que aprendi com Mário Covas foi a importância da sensibilid	facebook.com	2018-05-07T14:00:45+0000	07/05/2018 16:00	0	0	0	0	0	2400
Alckmi	photo	0	Um tema frequente nos encontros que tive no Estado do Rio de Janeiro foi a vi	facebook.com	2018-05-08T19:15:48+0000	08/05/2018 21:15	1	0	0	0	1	716
Bolson	video	0	UM SÓ BRASIL . Inscreva-se gratuitamente em nosso canal do youtube: https:/	facebook.com	2018-05-23T02:14:12+0000	23/05/2018 04:14	0	0	0	1	1	50935
Alckmi	video	6	Um prazer participar hoje da Agrishow em Ribeirão Preto grande feira do agr	facebook.com	2018-05-02T23:55:52+0000	03/05/2018 01:55	1	0	0	0	1	590
Alckmi	photo	6	Um dos grandes desafios do Brasil é também uma enorme oportunidade. Falta	facebook.com	2018-05-02T14:01:17+0000	02/05/2018 16:01	1	0	0	0	1	531
Alckmi	video	0	Um dia de muito trabalho e muito proveitoso em benefício do nosso país!	facebook.com	2018-05-31T01:05:42+0000	31/05/2018 03:05	1	1	0	0	2	1869
Ciro	photo	0	Turma quase completa. Faltou o caçula Gael!	facebook.com	2018-05-26T22:01:00+0000	27/05/2018 00:01	0	0	0	0	0	2196

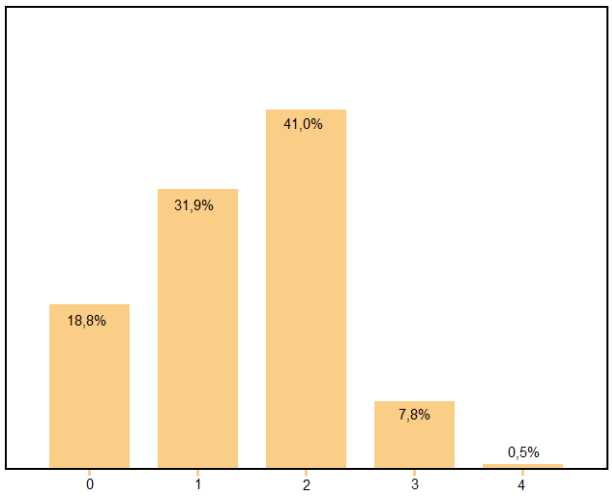
Antes da análise de classificações é preciso fazer testes de consistências, para identificar se a distribuição proposta pelo algoritmo de Reinert apresenta significância técnica. A primeira verificação é como as postagens foram classificadas nos *clusters*. Aqui, esperam-se duas coisas para haver consistência. A primeira é que a maior parte dos *posts* apresente pelo menos um dos termos que permita a classificação delas em um dos grupos. A segunda é se os *clusters* apresentam alguma distinção externa. Embora os termos não sejam excludentes, podendo aparecer em mais de um *cluster*, se as postagens forem classificadas em sua maioria em duas ou mais categorias, a classificação perde sentido.

O gráfico 4.2 a seguir mostra a distribuição das publicações por número de categorias que elas apresentam. A primeira informação importante é que apenas 18,8% das postagens não foram classificadas pelo método Reinert em nenhuma das categorias. O percentual acima de 80% de classificação dos textos como válidos mostra adequação inicial do método de classificação. Além disso, a maior parte das postagens está em um ou dois *clusters*. Apenas 7,8% fazem parte de três categorias e 0,5% em todas as quatro.

A tabela ao lado do gráfico 4.2 mostra duas informações complementares para o teste de confiabilidade. A primeira delas é a matriz de covariância entre os itens. Se a classificação estiver adequada, espera-se covariância positiva apenas entre os próprios *clusters* e negativas com os demais. Isso mostra que quando uma postagem está em um *cluster*, ela tende a não fazer parte de outro. A linha transversal, em cinza, mostra os coeficientes para o mesmo item. Em todos os casos ele é positivo e superior aos demais. Além disso, na maior parte das covariâncias com os demais *clusters* o coeficiente é negativo, excetuando a covariância entre CL2 e CL3. Como já vimos na tabela 4.1, as classes 2 e 3 são compostas por temos presentes predominantemente na *fanpage* de Lula. Isso é explicado pelo fato de se tratar de *clusters* com parte de termos comuns em textos da mesma fonte. São os *clusters* formados por postagens que tratam de Lula pré-candidato e de Lula réu na Lava Jato. O algoritmo de Reinert identificou diferenças entre os conjuntos de textos, porém, na prática eles são mais próximos entre si do que entre os demais grupos.

Por fim, a tabela também indica o coeficiente geral de teste de confiabilidade, o *α de Cronbach*. Neste caso, espera-se um coeficiente negativo, que indica que os *clusters* não variam em mesma direção, ou seja, não conformam componentes do mesmo indicador, que é o que se espera aqui – que os indicadores sejam independentes entre si. O coeficiente de -0,268 mostra que os termos de um *cluster* tendem a não estar em outro, como esperado aqui. Em outras palavras, quanto mais próximo de -1,000 for o coeficiente *α de Cronbach*, mais independentes serão os subconjuntos de textos presentes em cada uma das classes formadas pelo método de Reinert.

Gráfico 4.2. Distribuição dos temas e confiabilidade das categorias



Matriz de covariância Inter-item				
classe	CL1	CL2	CL3	CL4
CL1	0,185	-0,040	-0,037	-0,024
CL2	-0,040	0,251	0,236	-0,074
CL3	-0,037	0,236	0,250	-0,078
CL4	-0,024	-0,074	-0,078	0,153
α de Cronbach = -0,268				

Fonte: autor

Vencida a etapa de testes de consistência da classificação, é possível passar à seguinte, a da distribuição das categorias e, como se trata de textos em Redes Sociais Online, identificar o grau de interação e engajamento obtido pelas categorias. As seguintes perguntas podem ser feitas para direcionar as análises descritivas dos conteúdos: Como se dá a distribuição das categorias por *fanpage* de cada político? Como se dá a distribuição de engajamentos e interações a textos segundo as categorias a que pertencem? Existe diferença entre as interações alcançadas pelos pré-candidatos em função dos conteúdos das postagens?

Para responder à primeira pergunta, a tabela 4.2 a seguir sumariza os totais de postagens no mês de maio por pré-candidato, o número de publicações por página classificadas em pelo menos uma categoria e as ocorrências de postagens por tema em cada página. Os percentuais permitem comparar as distribuições entre as páginas, para identificar em que *fanpage* há maior percentual de ocorrências de determinado tema. Dos 374 *posts* que compõem o *corpus* de textos analisados aqui, 49,6% foram da *fanpage* de Lula, 20,1% de Alckmin, 15,9% de Ciro Gomes e 14,4% de Bolsonaro. No entanto, do total de postagens, 301 (80,7%) foram classificadas em pelo menos uma categoria temática. A tabela 4.2 a seguir permite identificar diferenças relevantes entre as páginas. A que teve menos conteúdo classificado foi a de Bolsonaro, com 42,5% de postagens classificadas, o que significa que mais da metade dos *posts* de Bolsonaro não puderam ser agrupados em uma das quatro categorias apresentadas pelo método

de Reinert. No outro extremo está a página de Lula, com 94,5% das postagens classificadas. Essas diferenças de percentuais inseridos no sistema de classificação é um indicador da concentração ou dispersão temática em cada página. No caso do político Bolsonaro, as postagens usam termos mais heterogêneos e dispersos, quando na de Lula há concentração terminológica, garantindo maior percentual de inclusão de textos nas categorias identificadas por intensidade de presença.

Tabela 4.2. Distribuição das postagens e classificação pelos temas nas *fanpages*

Página	Total Posts		Posts com tema		Emprego		Lula candidato		Lula réu		Campanha	
Alckmin	75	20,1%	61	81,3%	54	59,3%	6	3,2%	2	1,1%	8	11,4%
Bolsonaro	54	14,4%	23	42,5%	2	2,2%	3	1,6%			20	28,5%
Ciro	59	15,9%	42	71,2%	5	5,6%	2	1,2%	1	0,5%	38	54,3%
Lula	185	49,6%	175	94,5%	30	32,9%	173	94%	172	98,3%	4	5,8%
Total	373	100%	301	80,7%	91	100%	184	100%	175	100%	70	100%

Fonte: autor

O tema emprego teve maior concentração na *fanpage* de Alckmin, com 59,3% do total de postagens sobre esse assunto, ficando bem acima da participação do candidato do PSDB no total de *posts*. Depois vem a página de Lula, com 32,9%, porém, abaixo do percentual de participação dele no total de postagens. As páginas de Ciro e Bolsonaro tiveram participações mínimas no tema Emprego, com 5,6% e 2,2% respectivamente. Os temas Lula Candidato e Lula Réu, como previsto, estão concentrados na página de Lula, com 94% e 98,3% respectivamente. Já o tema "Campanha" predomina na página de Ciro Gomes, com 54,3% do total, seguida de Bolsonaro, 28,5%; Alckmin, 11,4%; e Lula, apenas 5,8%.

Como vemos, a página de Alckmin foi a que mais abordou o tema Economia, Lula tratou dele mesmo como candidatou ou réu, e Ciro Gomes e Bolsonaro usaram suas páginas para remeter a atividades de pré-campanha fora do Facebook. De maneira geral, eles não usaram nem para discutir tema de política pública, nem para promover suas próprias campanhas dentro da rede social online no período analisado. São comportamentos distintos identificados pela Análise de Conteúdo automatizada.

Feitas as descrições dos conteúdos das postagens, a próxima pergunta a ser respondida diz respeito ao impacto desses textos, ou seja, ao engajamento gerado por

eles. O Facebook considera engajamento de uma postagem a soma de comentários, reações e compartilhamentos do texto. Aqui, utilizamos o período mínimo de uma semana e máximo de quatro semanas após a data da publicação para coleta de dados sobre o engajamento aos *posts*. As quatro páginas receberam, juntas, no mês de maio 5,7 milhões de engajamentos. Desse total, 2,9 milhões na página de Lula (53,2%) e 2,3 milhões em Bolsonaro (40,4%). Ciro Gomes, com 250 mil (4,4%) e Alckmin com 178,8 mil (3,1%) ficam com os engajamentos totais baixos. No entanto, quando consideramos as médias de engajamentos por postagem, as posições superiores se invertem. Isso porque Bolsonaro tem quase o mesmo número de engajamentos que Lula, porém com bem menos *posts*. Assim, a média de engajamento por postagem de Bolsonaro fica em 42,8 mil, enquanto a de Lula está em 16,1 mil. Ciro Gomes tem 4,2 mil de média e Alckmin está na casa de 2,3 mil engajamentos de média por *post* (ver tab. 4.3 a seguir).

A estatística F (43,016) do teste de médias mostra que os totais de engajamentos têm diferenças estatisticamente significativas. A questão é saber se as distinções são entre todas as quatro páginas ou apenas entre algumas delas. O teste *Tukey post-hoc* cujos resultados também aparecem na tabela 3 mostra que os engajamentos apresentam três diferenças estatisticamente significativas. Os engajamentos médios em postagens de Alckmin e Ciro Gomes são os mais baixos e não têm diferenças estatísticas. Depois, segue o engajamento de Lula, em outro grupo, e acima de todos eles, com significância estatística, estão os engajamentos às postagens na *fanpage* de Bolsonaro.

Tabela 4.3. Médias e totais de engajamento por post e página em maio de 2018

Página	Post	%Post	Engajamento			
			Med./Post	%	Total	%
Alckmin	75	20,1	2.384	3,6	178.833	3,1
Bolsonaro	54	14,5	42.862	65,3	2.314.561	40,4
Ciro	59	15,8	4.244	6,5	250.371	4,4
Lula	185	49,6	16.137	24,6	2.985.408	52,1
Total	373	100	65.628	100	5.729.173	100

Teste de diferença de médias F = 43,016 (0,000)

Teste Tukey de Diferença de médias

Página	N	Subconjuntos por α = 0.05		
		1	2	3
Alckmin	75	2.384		
Ciro	59	4.244		
Lula	185		16.137	
Bolsonaro	54			42.862
Sig. (Tukey)		0,954	1	1

Fonte: autor

Agora que já conhecemos as interações gerais por página, o próximo passo é cruzar os conteúdos, a partir das categorias identificadas, com os engajamentos para identificar se determinado tema apresenta maior volume de interações do que outro. Como as postagens foram divididas em quatro classes temáticas, é possível imaginar que determinado tema possa gerar mais engajamento a um candidato do que a outros. Também é possível verificar se a concentração de determinado tema na *fanpage* de um candidato pode ser explicada pela maior interação obtida por ele nos *posts* relacionados ao tema. A tabela 4.4 a seguir apresenta testes de diferenças de médias univariados para o engajamento presente em postagens de cada uma das classes temáticas. Além das médias de interações individuais, as tabelas indicam o número de postagens (N) de cada candidato por classe, a média de engajamentos por candidato e o número de subconjuntos gerados em cada uma das temáticas.

As distinções já aparecem nas significâncias estatísticas. Enquanto a tabela 4.3 acima mostra que a página de Bolsonaro apresenta diferença estatisticamente significativa de todas as demais, com maior número de engajamento, quando consideradas as classes temáticas em separado, apenas na CL4 – "Campanha", Bolsonaro mantém a posição de maior engajamento com diferença estatisticamente significativa, pelo teste de *Tukey*. Nas outras três classes, as diferenças caem e ele divide o subconjunto 2 com a *fanpage* de Lula. No caso da CL1 – "Emprego", a média de Lula é de 14,03 mil engajamentos, acima dos de Bolsonaro, com 13,17 mil. Neste tema, Bolsonaro apresenta apenas um *post* no mês de maio. Nas classes CL4 - "Lula Candidato" e CL3 - "Lula Réu" Bolsonaro mantém o maior número de engajamentos, porém, com uma distância menor de Lula do que na CL4.

Tabela 4.4. Médias de engajamento por classe temática e *fanpage*

CL1 – EMPREGO			
Fanpage	**N**	**Subconjunto engajamento**	
		1	**2**
Alckmin	57	1.893,00	
Ciro	4	4.099,75	
Bolsonaro	1		13.171,00
Lula	31		14.034,90
Sig. (Tukey)		0,864	0,058

CL2 – LULA CANDIDATO			
Fanpage	**N**	**Subconjunto engajamento**	
		1	**2**
Alckmin	22	2.084,14	
Ciro	6	4.486,50	
Lula	178	16.292,80	16.292,80
Bolsonaro	3		29.806,67
Sig. (Tukey)		0,355	0,401

CL3 - LULA RÉU			
Fanpage	N	Subconjunto engajamento	
		1	2
Alckmin	7	2.229,00	
Ciro	4	8.487,00	
Lula	172	16.488,49	16.488,49
Bolsonaro	4		40.669,75
Sig. (Tukey)		0,506	0,088

CL4 – CAMPANHA			
Fanpage	N	Subconjunto engajamento	
		1	2
Alckmin	15	2.445,27	
Ciro	41	4.578,80	
Lula	7	16.934,00	
Bolsonaro	23		50.969,52
Sig. (Tukey)		0,656	1,000

Fonte: autor

A tabela 4.4 acima também permite comparar as médias de engajamento por candidato entre os temas. No caso de Alckmin, as médias são as mais baixas em todos os temas, além de apresentar baixa variação, indo de 1,8 mil no tema Emprego para 2,4 mil de média no tema campanha. É curioso verificar que a concentração de postagens na classe Emprego da *fanpage* de Alckmin não garante maior engajamento ao ele, ao contrário, é a menor média do candidato. No caso de Ciro Gomes, que tem o segundo menor engajamento, há variações em torno de quatro mil engajamentos em três categorias. Apenas em Lula Réu que o engajamento dele sobe para mais de oito mil interações. No caso de Lula, há manutenção em torno de 16 mil interações para três categorias e cai para 14,03 mil na classe Emprego. A *fanpage* de Bolsonaro é a que apresenta as maiores variações de médias entre as classes. Ele vai de 13,17 mil em Emprego, 29,80 mil em Lula Candidato, 40,66 mil em Lula Réu e 50,96 mil interações em Campanha.

Este capítulo apresentou uma técnica de Análise de Conteúdo automatizada para abordagem léxica de textos publicados em postagens ou comentários em Redes Sociais Online. No caso específico da análise desenvolvida aqui, foram incluídas apenas postagens em RSO e não os comentários. A proposta é substituir a adoção prévia de um livro de códigos de termos para indicação de temáticas pela formação de categorias temáticas a partir de *clusters* léxicos gerados pelos próprios textos. Dito de outra forma, o ensaio prático realizado aqui propõe que os textos passem por um tratamento inicial com o objetivo de identificar quais termos aparecem e como esses termos se distribuem ao longo do texto. Essa é a primeira etapa da análise automatizada. A segunda é, a partir das categorias extraídas dos próprios textos, codificar o conjunto textos para analisar frequências absolutas e relativas, além de comparar as características textuais

com variáveis externas, como indicadores de autoria, temporalidade ou – no caso de Redes Sociais Online – o engajamento que determinadas temáticas geraram.

Os resultados também mostraram que as categorias do exemplo apresentado aqui se distribuíram de maneira desigual entre as páginas. Como esperado, a página de Lula concentrou a maior parte de postagens com termos presentes nos *clusters* de Lula como pré-candidato e como réu. No entanto, em outras páginas esses *clusters* também apareceram, principalmente na de Bolsonaro. O *cluster* de economia concentrou-se na página de Alckmin, enquanto o de divulgação das atividades dividiu-se entre as páginas de Bolsonaro e Ciro Gomes. Ou seja, além de os textos formarem apenas um agrupamento temático, a classificação posterior das postagens indicou que esse *cluster* está concentrado na página de um dos quatro políticos inseridos na análise. Por se tratar de textos em páginas de rede social online, existem informações sobre o impacto de cada postagem, via interações, comentários e compartilhamentos, o chamado engajamento ao conteúdo. Apesar das limitações da técnica, o exemplo apresentado aqui se mostrou mais adequado do que o uso do livro de códigos para textos em RSO.

4.7 Referências bibliográficas do Capítulo IV

Alonso, S., *et al.* (2012). *Análisis de contenido de textos políticos*. Un enfoque cuantitativo. Madrid: Centro de investigaciones Sociológicas.

Bardin, L. (1977). *Análise de Conteúdo*. Lisboa: Edições 70.

Benzécri, J. P. (1992). *Correspondence analysis handbook*. New York: Marcel Dekker.

Berelson, B. (1984). *Content analysis in communication research*. New York: Hafner.

Camargo, B. V., & Justo, A. M. (2013). IRAMUTEQ: um software gratuito para análise de dados textuais. *Temas em Psicologia*, 21(2), 513-518. Disponível em: http://www.redalyc.org/pdf/5137/513751532016.pdf. Acesso em junho de 2018.

Drisko, J. (2016). *Content Analysis*. Oxford: Oxford University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: the promise and pitfalls of automatic content *analysis* methods for political texts. *Political Analysis*, 21(3), 267-297.

Krippendorff, K. (2004). *Content Analysis: An introduction to its methodology*. Thousand

Oaks, CA: Sage Publications.

Lombard, M. (2004). A call for standardization in content *analysis* reliability. *Human Communication Research*, 30(3), 434-437.

Neuendorf, K. A. (2016). *The content analysis guidebook*. Thousand Oaks, CA: Sage Publications.

Reinert, M. (1990). Alceste, une méthodologie d'analyse des données textuelles et une application: Aurelia de Gerard de Nerval. *Bulletin de Methodologie Sociologique*, 26, 24-54.

Rieder, B. (2013). Studying Facebook via data extraction. In *Proceedings of the 5th annual ACM web science conference*. ACM, WebSci´13, Paris, France, 346-355. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.6806&rep=rep1&type=pdf. Acesso em julho de 2018.

Riffe, D., Lacy, S. & Fico, F. (2014). *Analyzing media messages: using quantitative content analysis in research*. New York: Routlege Taylor & Francis Group.

White, R. K. (1947). Black boy: a value-analysis. *The Journal of Abnormal and Social Psychology*, 42(4), 440-461.

4.8 Exercícios propostos do Capítulo IV

4.8.1 A partir do banco de dados "BDCap4AC" disponível em https://blogempublico.files.wordpress.com/2018/09/bdcap4ac.docx, insira o *corpus* textual no Iramuteq. Nele, rode um teste de Reinert e identifique os *clusters* formados a partir dos termos presentes nas postagens das páginas oficiais dos diretórios nacionais de três partidos políticos brasileiros no mês de junho de 2018 (Patriotas, PSDB e Rede). A partir dos resultados, monte um quadro indicando os termos com significância estatística para cada classe, os percentuais de ocorrências, de postagens e de casos, a exemplo do quadro 4.1 do capítulo.

4.8.2 Use o a planilha "BDCAP4AC_tab" disponível em https://blogempublico.files.wordpress.com/2018/09/bdcap4ac_tab.xlsx para preencher os dados das classes. Use as últimas colunas à direita da planilha com 0 = não consta termo do *cluster*, 1 = consta termo do *cluster* para o total de postagens do banco de dados nas colunas CL1, CL2, CL3, CL4, CL5 e CL6. (Sugiro o uso dos filtros na planilha de dados para identificar os termos e preencher as colunas)

Capítulo V

Análise de Redes Sociais (ARS)

Redes sociais oferecem dados detalhados sobre processos sociais não capturados diretamente nos indivíduos, grupos ou instituições, mas em suas conexões.

Este capítulo é dedicado à apresentação da técnica de análise de redes sociais. Aplica-se este tipo de análise quando o objetivo é pesquisar não os indivíduos, seus coletivos, suas instituições ou resultados de suas ações isoladamente. Aqui o que se busca é explicar as conexões entre indivíduos, instituições ou indivíduos e instituições para descrever a forma como se dão as relações entre eles, ou seja, queremos conhecer como são formadas as redes sociais e não os resultados diretos das ações daqueles que integram as redes. Aqui, usaremos para o exemplo prático um *plug-in* do Excel para a geração de gráficos e estatísticas de rede chamado NodeXL[1]. Existem outros pacotes e *softwares* de código aberto que desenvolvem esse tipo de análise. Além disso, outro aviso importante é que o capítulo apresenta apenas as aplicações mais bá-

[1] Existem diferentes *softwares* que atualmente fazem análises de redes sociais, inclusive muitos deles no formato *open source*, sem a necessidade de pagamento de licenças para uso. Neste tópico as redes serão calculadas pelo *software* NODEXL, desenvolvido por um consórcio de pesquisadores de cincos universidades: Cambridge, Maryland, Stanford, Porto e Oxford. Hoje, ele é mantido por uma entidade sem fins lucrativos, a *Social Media Research Foundation*. Pode ser obtido gratuitamente no endereço eletrônico (http://nodexl.codeplex.com/releases/view/117659), assim como manuais para utilização do *software*.

sicas da análise de redes, para iniciantes. O interessado nesse tipo de técnica deve se aprofundar em conceitos como centralidade, intermediação, densidade e posição periférica, que não poderão ser detalhados aqui. Os exemplos, tanto do capítulo, quanto dos exercícios, são sobre redes de partidos políticos, características de seus candidatos e tipos de doadores para suas campanhas. No entanto, mais recentemente as aplicações de análises de rede têm se desenvolvido principalmente em pesquisas sobre interações em Redes Sociais Online (RSO).

5.1 Conceituando Análise de Redes Sociais (ARS)

A análise de Redes Sociais (ARS) é uma técnica específica usada em pesquisas empíricas para medir níveis de interações entre os atores sociais com muitas aplicações no campo da ciência política, principalmente depois do surgimento das Redes Sociais Online (RSO). Existem várias formas de conceituar uma rede social. Para Bourdieu (2003), uma rede é uma configuração de relações objetivas entre posições, definidas tanto pela sua própria existência quanto pelas determinações dos ocupantes dessas posições. Segundo Burt (1984), redes sociais são um conjunto de atores conectados por relações sociais específicas. Portanto, nas análises de redes sociais o interesse maior está nas ligações e papeis desempenhados pelos integrantes das redes nas interações e não nos atores propriamente ditos. Esse tipo de abordagem fundamenta-se no fato de que os atores políticos são interdependentes e que isso traz consequências relevantes para cada integrante de uma rede (Freeman, 1979). O que interessa aqui como ponto de partida é o posicionamento na estrutura de uma rede em relação aos demais atores e não o ator em si.

A análise de redes sociais dá prioridade às relações entre os atores envolvidos nos processos políticos, diferenciando-se das técnicas que visam descrever as características próprias dos atores (abordagem microssociológica) ou das estruturas organizacionais e sociais que limitam e constrangem as ações individuais (abordagem macrossociológica). O ponto forte da técnica de ARS é permitir uma superação da dicotomia Micro X Macro. Nela, o objetivo é estudar como os atores políticos (sejam eles indivíduos ou instituições) se organizam relacionalmente tanto entre si quanto em um

ambiente maior.

Como defende Marques (2007), não é possível pensar em relações sociais apenas considerando as características individuais de cada ator envolvido na relação ou analisar de forma abstrata o ambiente institucional sob o qual as relações acontecem. É preciso levar em conta os diferentes mecanismos relacionais, o que significa considerar os pesos das instituições e das decisões individuais nos posicionamentos dos atores políticos no espaço relacional. Por outro lado, deve-se reconhecer as limitações das ARS como técnica de análise empírica: ao permitir uma identificação relacional, o uso dessa técnica é limitada quanto às explicações sobre a natureza dos atores individuais ou sobre a composição mais geral das instituições que fazem parte das estruturas relacionais. Em outras palavras, não é possível fazer inferências sobre as intenções individuais dos atores envolvidos nas relações sociais a partir da ARS, assim como também não se pode pensar em explicações sobre a natureza e a origem das organizações que se relacionam. A técnica não foi pensada para isso. Ela serve para medir as interações entre os atores, suas intensidades, direções e força das relações. Portanto, na ARS, a unidade de análise é a relação e não os atores envolvidos ou as organizações isoladamente.

Hanneman e Riddle (2005) lembram que a ARS possui uma linguagem própria para descrever a estrutura e o conteúdo das relações observadas, fugindo das preocupações sobre quão fortes ou fracos, iguais ou desiguais, são os atores envolvidos. Ela centra atenção em como se localizam os atores envolvidos nas relações. Para Costa (2011) há quatro aspectos importantes a serem considerados em ARS. O primeiro deles é que a técnica não desconsidera as características e atributos dos indivíduos envolvidos nas relações. Ao contrário, esses atributos são considerados em suas proporções como possíveis explicações para o tipo de relação encontrada, ainda que identificar atributos dos atores não seja o objetivo final na ARS. Em segundo lugar, a análise de redes sociais é uma metodologia que depende diretamente das relações entre conceitos teóricos e dados empíricos, sendo mais que uma simples técnica empírica. Sem consistência conceitual sobre quem são e quais os interesses dos atores envolvidos, as explicações sobre as relações não se sustentam. O terceiro é que se trata de uma análise estrutural, das estruturas de relações. Para tanto, seu sucesso depende de rigor metodológico, empírico e matemático para que os “achados” sobre as estruturas das

relações sejam plausíveis. Em quarto, a metáfora da rede exige a apropriação de determinados conceitos necessários para diferenciá-la de outros estudos que usam o termo "rede", porém têm o objetivo de analisar os atores e não suas relações (Costa, 2011). A seguir, são sumarizados os principais conceitos aplicados às análises de redes sociais.

5.2 Componentes da ARS

Por definição, rede é um conjunto de laços do mesmo tipo entre dois ou mais atores, que podem ser indivíduos ou instituições. Os laços são episódicos em uma relação social, portanto, a ARS é uma técnica para análise relacional e episódica. Os principais conceitos dessa técnica são:

- Nós: são atores envolvidos nas relações, podem ser indivíduos, instituições, países, etc.

- Arestas, laços ou arcos: representam as ligações/relações entre os nós de uma rede.

- Atributos: são as características dos nós que interessam ao pesquisador. São exemplos de atributos: status social, nível educacional, sexo, ocupação, região do país, partido político, etc. Os atributos são representados por cores no mapa de rede.

- Relação unidirecional (arestas): indica que a relação começa em um nó e toma uma única direção, sem que haja reciprocidade na relacional.

- Relação bidirecional (arcos): indica relações mútuas entre os nós ligados pela aresta, que nesse caso recebe o nome de arco.

- Transitividade (intermediação): é uma característica das relações/laços. Uma relação transitiva é aquela que permite a conexão entre dois atores separados na rede, porém conectados a um terceiro ator. Por exemplo, se X tiver conectado com W e se T tiver conexão com W diz-se que W é transitivo entre X e T. A transitividade gera uma tendência de agrupamentos que subdividem as redes por características similares. Os grupos formados por essas subdivisões são chamados de *clusters*.

- Coeficiente de *Clustering*: indica a capacidade natural de um laço transmitir relações entre nós. Um coeficiente alto mostra relações muito densas entre os nós da

rede, com todos mais ou menos se relacionando com os demais. Um coeficiente baixo mostra pouca transitividade de relações ou relações não transitivas.

- Densidade: é medida pela razão entre o número de ligações de um grafo com o número máximo de ligações caso todos os vértices estivessem conectados entre si. No limite, a densidade máxima é Um (1), que indicaria que todos os nós estão conectados entre si, mostrando existir altos níveis de interação; enquanto a densidade mínima, próxima de zero, indica que há um número muito pequeno de conexões efetivas em relação ao potencial total.

- Modularidade: é um indicador da homogeneidade da distribuição de nós e vértices na rede. Quanto mais equidistantes forem as extremidades de uma rede, mais modular ela será. Indica possíveis ocorrências de nós fracos (distantes) do centro da rede.

- Diâmetro: é um coeficiente que indica as características gerais da rede. Quanto maior o diâmetro, mais ampla será a rede. Interessante notar que ao acrescentar um nó, não necessariamente o diâmetro aumenta. O diâmetro da rede só cresce quando surge um novo nó periférico. Se não estiver na periferia das relações, o novo nó apenas adensará a rede sem aumentar seu diâmetro.

- Distância geodésica média: é o coeficiente que indica a média da distância entre dois nós quaisquer na rede. Por exemplo, uma distância média de 1,52 indica que são necessárias mais de uma e menos de duas conexões, na média, para um nó acessar outro qualquer na rede, tomando o caminho mais curto. Por considerar o caminho mais curto é que recebe o nome de geodésica.

- *Closeness* (proximidade): medida de centralidade que identifica quão próximo está um ator/nó de todos os demais na estrutura da rede. Um nó é considerado central se ele pode interagir com todos os demais rapidamente. O coeficiente é obtido pelo cálculo do inverso da soma dos menores caminhos para todos os nós.

- *Betweeness* (centralidade): medida de centralidade que indica onde se encontra um ator/nó no caminho mais curto entre dois outros nós quaisquer na estrutura da rede. Ele mostra o grau de interação entre os outros atores. É calculado pela soma das frações dos laços mais curtos de um nó em relação aos demais atores. Quanto mais central for o nó, mais ativo ele será na rede.

- **Matriz de contato:** é a matriz que representa as relações de um nó com todos

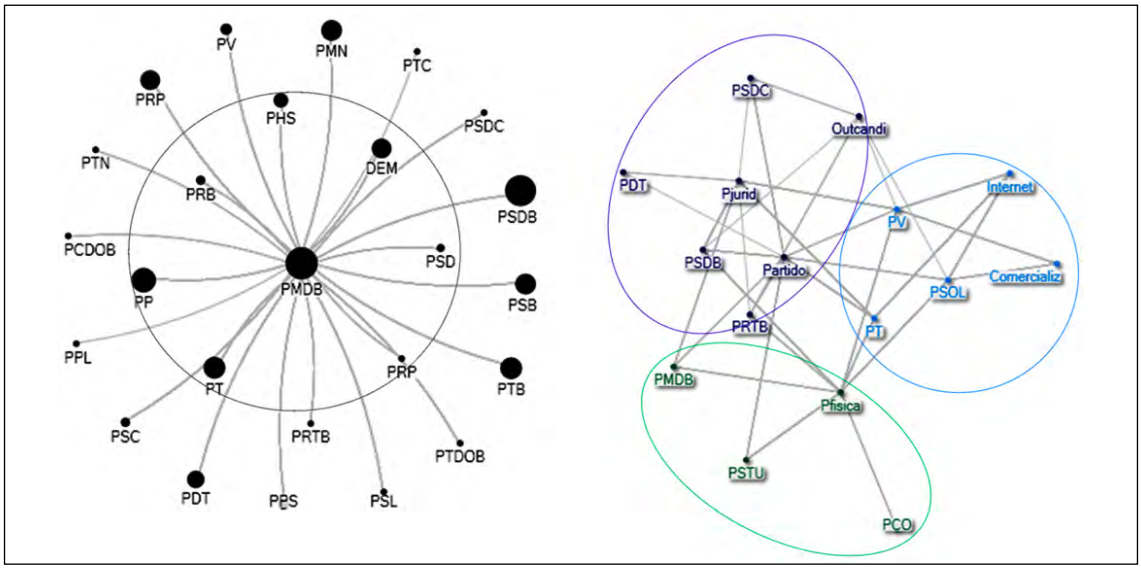
os demais. A representação numérica é binária. Se existe relação entre dois nós, ela é representada pelo número Um (1). Se não existe, a representação é com zero (0). Por exemplo: Matriz de contato do nó X = 1; 0; 0; 1. Em forma de matriz, as relações entre os nós X, Y, Z, T e W seria a que segue no quadro 5.1 abaixo (é evidente que só é necessária a leitura de uma das metades longitudinais da matriz).

Quadro 5.1. Exemplo de matriz de contato entre nós de uma ARS

	X	Y	Z	T	W
X	-				
Y	1	-			
Z	0	0	-		
T	0	1	1	-	
W	1	1	1	1	-

O foco de atenção da ARS nem sempre precisa estar nas relações predominantes ou naqueles nós centrais, como tende a acontecer em outras técnicas que buscam identificar padrões gerais de comportamentos. Ao contrário, a identificação de uma relação secundária, de um vértice fraco ou de um nó periférico pode ser mais explicativa do que as relações principais. Existem dois tipos principais redes. As sociocêntricas, quando não há um centro definido para as relações, e as egocêntricas, que são organizadas ao redor de um nó. A seguir, o Grafo 5.1 apresenta um exemplo para cada um dos dois tipos de redes.

Grafo 5.1. Exemplos de redes ego e sociocentradas de coligações partidárias



Fonte: autor a partir de dados do TSE.

A imagem da esquerda é um exemplo de rede **egocentrada** de coligações de partidos em apoio a candidatos a prefeito do PMDB em pequenos municípios (até 10 mil eleitores) em 2012. Há um conjunto de nós mais próximos e outro mais distante. Como a rede é formada pelo número de municípios em que houve coligação entre PMDB e demais partidos, a maior proximidade indicada uma relação mais intensa, ou seja, maior número de coligações municipais entre PMDB e aqueles partidos. Por outro lado, as siglas mais periféricas apresentaram um menor número de coligações municipais com o PMDB naquele ano.

A imagem da direita é um exemplo de rede **sociocentrada** com três *clusters* entre número de doações recebidas (e não volume de recursos recebidos) por tipo de doador e sigla do partido - para diretório nacional em 2010. Os doadores podem ser Pessoa Física, Pessoa Jurídica, Outro Candidato, o próprio Partido, resultado de Comercialização de produtos ou doações por Internet. Quanto maior o número de operações de um tipo de doação com um partido, maior a chance de formação de um *cluster* entre esses dois nós. No caso, formou-se um *cluster* entre Doações de Pessoas Físicas e PMDB, PSTU e PCO, indicando maior número de interações desses partidos com o tipo de doação por pessoa física, quando comparado aos demais. As doações de pessoas jurídicas, dos próprios partidos e de outros candidatos estão no *cluster* com PSDB, PDT, PSDC e PRTB. Já o PV, PT e PSOL integram um grupo com mais operações de doações por Internet e por comercialização de produtos. Perceba que, para além dos *clusters*, em cores distintas e circunscritos a seus círculos, a imagem mostra as arestas de ligação entre partidos e todos os tipos de doações que receberam em 2010, ainda que de forma menos intensa.

Cada novo laço na rede acrescenta valor aos indivíduos ou grupos conectados. No caso do grafo 5.1 acima, as relações entre eles permitem acesso a recursos, portanto, quanto mais conexões, melhor. Isso porque redes são estruturas através das quais passam recursos e é bom ter um número maior de laços porque eles dão acesso a bens sociais. As redes podem adotar diferentes estruturas e formatos. Esses formatos levam a distintas implicações para os nós. Quando os nós formam grupos subdivididos (formando *clusters*), eles tendem a não apresentar desempenho individual tão bom quanto teriam nos casos de redes com maior coesão.

Se a ARS existe para verificar a forma e intensidade de ligação entre nós co-

nectados de alguma forma, seu objetivo principal é a análise de relações entre atores sociais. Para tanto, ela parte do princípio de que os atores envolvidos criam ligações entre si para acessar recursos sociais e podem formar *clusters* quando o acesso aos recursos apresenta algum padrão, ou seja, conexões em rede ajudam a chegar até esses recursos. Os padrões de conexões formam imagens que são desenhadas e permitem identificar o que está acontecendo no mundo empírico das relações sociais, além de permitir identificar atores-chave e os que têm participações fracas nas redes.

5.3 Etapas para Análise de Redes Sociais

Existem cinco passos fundamentais apresentados por Freeman (1979) para uma correta construção de imagens de redes sociais:

i) desenho de gráficos,

ii) desenho de gráficos dirigidos,

iii) uso de cores para desenhar gráficos múltiplos;

iv) formatação dos nós para indicar os atributos dos atores sociais e

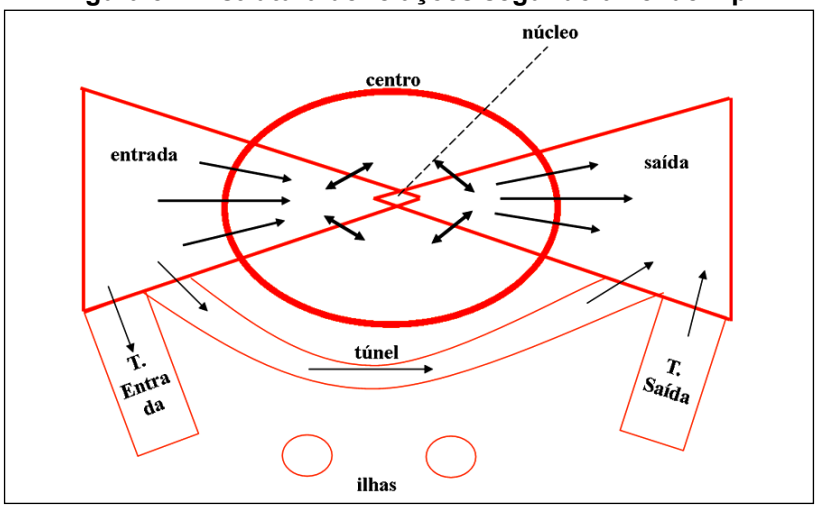
v) localização dos pontos na rede para indicar as propriedades estruturas dos dados.

Quando passa por essas etapas, o grafo produzido cumpre a função de fornecer informações visuais e estatísticas para a representação de realidades complexas com clareza e eficiência. Se estivermos analisando relações entre duas variáveis quaisquer, uma lei que ajuda a entender a associação entre elas é a “Lei de Potências”. De acordo com essa lei, quando a relação entre as variáveis se dá em escala logarítmica, uma linha reta é a melhor forma de descrever o tipo de relação.

Em ARS, uma variável é a quantidade de vezes que um ator aparece, chama-se de “Lei de Zipf” e ajuda a entender as dinâmicas de muitas redes do que diz respeito ao número e forma de conexões. Originalmente proposta pelo linguista George Kingsley Zipf como lei de potências sobre distribuição de valores em uma ordem qualquer em uma lista. Para a ARS, esse tipo de relação ocorre quando as conexões de nós mais próximos formam um múltiplo fixo do segundo grupo em termos de conexões e assim

por diante. Os nós mais conectados são aqueles mais conhecidos. A aplicação do modelo da “Lei de Zipf” para ARS está representada na figura a seguir.

Figura 5.1. Estrutura de relações segundo a Lei de Zipf

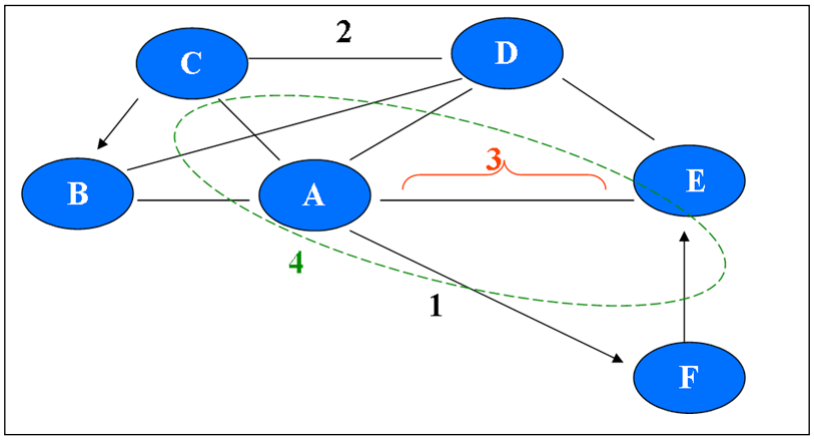


Fonte: autor a partir de Zipf, 1940.

Na ilustração acima, o gráfico dirigido é formado por um componente principal (centro) e duas asas. Uma de entrada e outra de saída, definindo a força das relações. Os componentes com relações mais fracas estão fora do componente principal e são Tentáculos (de entrada e de saída), além de ilhas, que são os componentes mais isolados da rede. Para uma aplicação do modelo, a figura 5.2 a seguir representa os principais elementos de uma ARS em uma aproximação com a Lei de Zipf. Os nós são os pontos de convergência (A, B, C, D, E, F). No caso, a rede é formada por seis nós, que podem ser indivíduos, organizações, instituições, partidos políticos, unidades geográficas como municípios ou países, etc. Os nós são ligados por laços ou vértices, que podem ser unidirecionais, quando a relação só tem uma direção e são indicados por uma seta. É o caso do vértice (1), que indica uma relação em direção única de A para F. Ou podem ser vértices bidirecionais, quando a direção da relação entre os nós é indefinida. No exemplo, isso é representado pelo vértice (2), ligando C a D. A centralidade é definida pelo número de vértices ligados a um nó. Quanto maior o número de vértices, mais central é o nó. Nesse caso, o nó (A) é o mais central por apresentar o maior número de ligações com outros nós. A proximidade entre os nós (3) indica a distância entre cada integrante da relação. Quanto mais distante, menor a força da relação. A Distância

Geodésica Média indica qual o grau de proximidade ou distanciamento entre os nós em uma rede. Quanto mais próximos os nós estiverem, maior a densidade da rede, o que indica maior integração entre seus componentes.

Figura 5.2. Exemplo de componentes que integram uma rede



Fonte: autor

O conceito de intermediação aplica-se aos casos em que há uma relação indireta entre os integrantes de determinada rede. No exemplo acima é possível pensar que o nó A faz intermediação entre os nós B e F, visto que não existe vértice ligando os dois últimos. B e F não apresentam uma relação direta. No máximo têm uma relação intermediada por fazerem parte da mesma rede social. A modularidade é um indicador que mostra quão homogênea é a distribuição dos nós e vértices em uma rede. No exemplo acima se vê pelo formato da rede (4) que suas extremidades não são equidistantes. Há uma distribuição maior do lado direito da rede e uma concentração do lado esquerdo. Esse formato reduz o valor da modularidade, indicando relações heterogêneas entre os integrantes da rede. Neste caso, a modularidade é distorcida pela presença do "nó fraco" F, que está mais distante do centro e apresenta o menor número de vértices. Por outro lado, os nós B e C apresentam maior proximidade do centro e homogeneidade de distâncias entre si que dos demais nós. Esses são considerados os "nós fortes".

Apresentadas as características básicas da ARS, vamos exemplificar a aplicação desse tipo de análise a um caso empírico. Usaremos aqui as informações sobre doações de empresas (pessoas jurídicas) para os diretórios nacionais dos partidos

políticos durante a campanha presidencial de 2010[2]. Nosso objetivo é montar a rede de distribuição e conexões das empresas doadoras e dos partidos que receberam as doações. Nessa rede são consideradas as informações de prestação de contas dos partidos políticos disponíveis no site do Tribunal Superior Eleitoral (TSE). Serão computadas apenas as doações feitas aos diretórios dos partidos, não aquelas registradas diretamente pelo candidato ou as computadas ao comitê financeiro das campanhas. Com isso, no próximo tópico, esperamos identificar a rede formalizada de doações financeiras entre empresas e partidos durante a campanha eleitoral e não entre empresas e candidatos.

5.4. Redes de financiamento de empresas a partidos políticos no Brasil

Seguindo o princípio basilar deste manual, optamos por utilizar aqui a forma mais simples de se produzir gráficos e análises de redes sociais com aplicativos de código aberto, os chamados *softwares* livres. Para o iniciante em análise de redes a sugestão é começar pelo NodeXL, desenvolvido pela *Social Media Research Foundation*, uma organização sem fins lucrativos destinada a criar ferramentas de análise de dados abertas. O NodeXL é um *plug-in* bastante simples da planilha de dados Excel. Uma vez instalado no computador, o *plug-in* abre em uma planilha, onde são realizadas todas as operações[3]. Para quem não conhece NodeXL, é necessário uma consulta a textos e tutoriais que ensinam o básico das operações antes de continuar neste capítulo. Pode-se consultar artigo da *Pew Research Foundation*[4] sobre o funcionamento básico do *plug-in* ou o artigo de Smith *et al.* (2009) que apresenta as funções do NodeXL para análise de redes sociais online[5]. Além do NodeXL, outros aplicativos de código aberto específicos para análise de rede são Pajek, Ucinete e Gephi, além de pacotes específicos do R para geração de grá-

[2] Até 2014 foi possível às empresas e demais pessoas jurídicas fazerem doações aos partidos e seus diretórios durante as campanhas eleitorais. Desde a minirreforma eleitoral de 2015, as doações de empresa foram proibidas.

[3] Disponível em: https://archive.codeplex.com/?p=nodexl

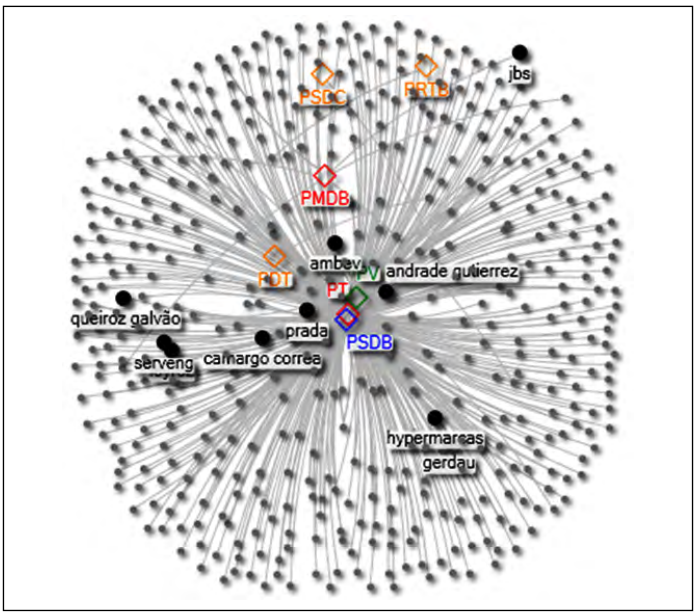
[4] Disponível em: http://www.pewinternet.org/files/2014/02/How-we-analyzed-Twitter-social-media-networks.pdf

[5] Disponível em: http://www.cs.umd.edu/hcil/trs/2009-11/2009-11.pdf

ficos de rede. Considerando que você já baixou o NodeXL, leu os manuais e tem algum conhecimento básico do funcionamento dele, vamos adiante com o exemplo.

Na eleição nacional de 2010, o TSE registrou 650 operações de doações de empresas a diretórios nacionais de partidos políticos durante a campanha. No entanto, no exemplo apresentado aqui, o gráfico apresenta apenas 484 vértices (ligações entre dois nós específicos). Isso significa que existem empresas que fizeram mais de uma doação a um mesmo partido ou a mais de um partido várias vezes. A rede formada entre doadores e partidos está representada nos gráficos a seguir. Depois, são apresentadas outras duas versões do mesmo gráfico, porém, com a indicação de *clusters* formados por métodos distintos. Além dos nomes dos partidos, o gráfico também mostra os maiores doadores (aqueles que fizeram doações acima de R$ 1 milhão por operação). No grafo 5.2 abaixo é possível perceber que existem três partidos centrais (PT, PSDB e PV). Os doadores mais centrais são Andrade Gutierrez, Ambev e Prada, ou seja, são aqueles que além de doarem altos valores, fizeram doações para mais partidos. A empresa JBS é grande doadora, porém apresenta laços fracos, pois está na periferia da rede. Isso é explicado pelo fato de a JBS ter feito doação de alto valor para apenas um partido político em 2010, o diretório nacional do PMDB. Da mesma forma, por terem recebido altas doações de poucas empresas, os partidos PRTB e PSDC também se encontram em áreas periféricas do grafo.

Grafo 5.2. Rede simples de maiores doadores a diretórios nacionais em 2010



Fonte: autor a partir do TSE.

A tabela 5.1 abaixo sumariza as principais estatísticas da rede. O número de vértices indica quantas conexões existem ao todo. Como temos 650 nós e 484 vértices, isso significa que alguns nós produziram mais de uma conexão, ou seja, fizeram doações com altos valores a mais de um diretório nacional. A distância geodésica média de 2,794 mostra que em média um doador está entre duas e três conexões de outro. A densidade é muito baixa, 0,002, indicando que a rede não explora todas as conexões potenciais. Isso já seria esperado, pois a máxima densidade aqui significaria que todas as empresas estão fazendo doações a todos os partidos políticos. A modularidade de 0,439 mostra certa homogeneidade nas distribuições/tamanhos dos laços em toda a rede. De qualquer maneira, podemos perceber que há três partidos centrais, a proporção de conexões reais em relação ao total potencial (densidade) é muito baixa e os atores distribuem-se de maneira homogênea na rede.

Tabela 5.1. Principais estatísticas da rede de maiores doadores a diretórios nacionais em 2010

Estatística	Valor
Vértices	484
Total de nós	650
Distância geodésica média	2,794
Densidade	0,002
Modularidade	0,439

Os indicadores da rede são mais úteis quando podem ser comparados. Por exemplo, na tabela 5.2, se considerássemos apenas a rede egocentrada do diretório nacional do PT (desconsiderando todos os demais partidos), para *clusters* individuais, teríamos uma distância geodésica média de 1,987 e densidade de 0,003. Ou seja, maior proximidade entre os nós quando comparado com o gráfico geral e um pouco mais de densidade. Os coeficientes para a rede do PSDB ficam próximos dos do PT, com distância geodésica média de 1,974 e densidade de 0,006. Para o PV, a distância geodésica média cai para 1,860 e a densidade sobe, passando para 0,036. Isso significa que para os diretórios nacionais dos três partidos, cada doador pode se conectar ao outro com menos de duas conexões, mas a densidade/proximidade entre os nós da rede do diretório nacional do PV foi bem superior à dos outros dois partidos. As maiores densidades e menores distâncias geodésicas ficam com PDT, PRTB e PSDC, no entanto, eles não podem ser comparados com os partidos anteriores porque o número vértices desses últimos é pequeno.

Tabela 5.2. *Clusters* por algoritmo de vizinhança

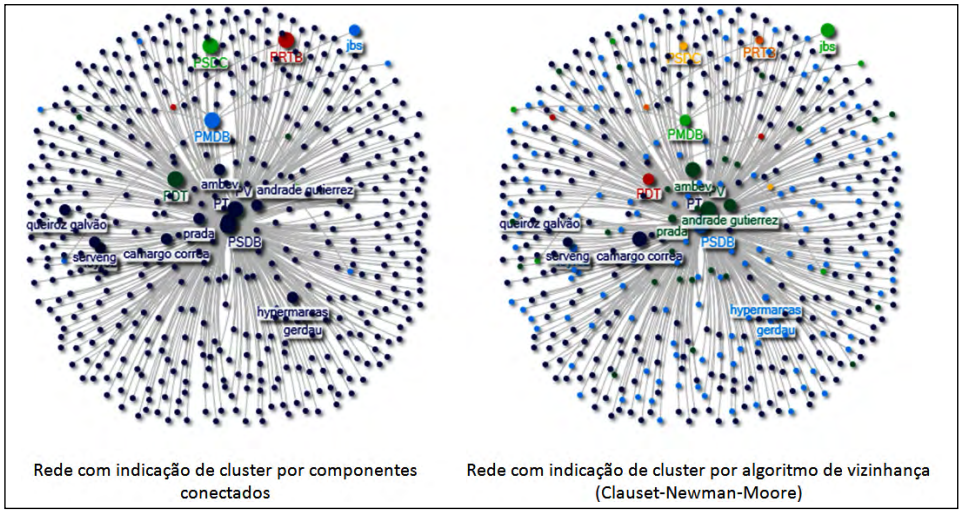
Clusters	Partido	Vértices	Distância Geodésica Média	Densidade
G1	**PT**	301	1,987	0,003
G2	**PSDB**	141	1,972	0,007
G3	**PV**	28	1,860	0,036
G4	**PMDB**	7	1,469	0,143
G5	**PDT**	3	0,889	0,333
G6	**PRTB**	2	0,500	0,500
G7	**PSDC**	2	0,500	0,500

Fonte: autor a partir de TSE.

Outra forma de analisar as relações em uma rede sociocêntrica (com vários partidos) é a partir dos grupos que se formam das conexões. Um desdobramento da ARS é a verificação de como os nós se organizam dentro da rede formada pelos *clusters* com características específicas. No entanto, é preciso saber que há diferentes formas de agrupar os nós em uma rede. Dependendo da maneira como se dá a montagem dos *clusters*, podemos ter respostas muito diferentes.

Os dois gráficos a seguir são exemplos dessas diferenças. O primeiro é a formação de *clusters* por componentes conectados, ou seja, independente de ser partido ou empresa, se houver conexão entre eles (direta ou indireta), eles tenderão a fazer parte do mesmo grupo. O segundo é uma formação de *clusters* por algoritmo de vizinhança. Significa que nós/empresas próximos a cada um dos nós/partidos tenderão a formar um *cluster*. As cores indicam as composições dos grupos.

Grafo 5.3. Comparação entre dois métodos de formação de *clusters*



Fonte: autor a partir de TSE.

Como podemos perceber no grafo 5.3 acima, há um número menor de *clusters* por componentes conectados do que pelo algoritmo de vizinhança. São cinco *clusters* por componentes conectados e sete por vizinhança. Isso porque, no primeiro, os três partidos centrais (PT, PSDB e PV) fazem parte do mesmo grupo, enquanto no segundo cada um organiza um *cluster* específico. A diferença é explicada pelo fato de que várias empresas doam para mais de um dos três partidos e, portanto, isso os conecta um mesmo grupo, enquanto que no *cluster* por vizinhança esse efeito de transitividade é desconsiderado. Por este motivo, na tabela 5.2, para analisar as densidades e distâncias geodésicas dos *clusters* dos partidos, utilizamos o método de algoritmo de vizinhança para análises de redes egocentradas em partidos únicos. A tabela 5.3 a seguir sumariza os principais coeficientes dos *clusters* por componentes conectados.

Tabela 5.3. *Clusters* por componentes conectados

Clusters	Partido	Vértices	Distância Geodésica Média	Densidade
G1 (azul escuro)	**PT, PSDB, PV**	470	2,795	0,002
G2 (azul claro)	**PMDB**	7	1,469	0,143
G3 (verde escuro)	**PDT**	3	0,889	0,333
G4 (verde claro)	**PSDC**	2	0,500	0,500
G5 (vermelho)	**PRTB**	2	0,500	0,500

Fonte: autor a partir de TSE.

A tabela 5.3 acima descreve as principais estatísticas da rede, agora divididas pelo método de componentes conectados. Percebe-se que o primeiro *cluster* tem o maior número de vértices, maior distância geodésica e menor densidade, como esperado. A densidade vai aumentando conforme cai o número de vértices e a distância geodésica. Nota-se uma queda mais significativa entre o G1 e G2 do que entre G2 e G3.

A rede com indicação de *clusters* por vizinhanças permite identificar outro padrão de conexões, como indicado na tabela 5.2 acima. O maior *cluster* (1) é formado pelos doadores ao diretório nacional do PT, com 301 vértices. Já nos grafos, é possível perceber que os maiores doadores ao diretório nacional do PT foram Camargo Correia, Serveng e Queiróz Galvão, notadamente do setor de construção civil. O segundo *cluster* é formado por doadores do PSDB, com 141 vértices e os principais doadores em volume de recursos são Hypermarcas e Gerdau. O terceiro *cluster* é do PV, com apenas

28 vértices, destacam-se Ambev, Camargo Correia e Prada. Importante notar que o número de nós em um *cluster* não tem relação com o volume de recursos que o partido recebe, mas sim com o número de doadores diferentes. O *cluster* 4 é do PMDB, contando com apenas 6 nós, onde se destaca a empresa JBS. O PDT tem um *cluster* com apenas dois vértices e, por fim, temos PRTB e PSDC com apenas um vértice cada um entre as empresas doadoras de recursos a seus diretórios nacionais. Como na tabela 5.2, os grupos estão formados por vizinhança de doadores a partidos. As comparações entre os *clusters* permitem concluir que o diretório nacional do PT apresentou a mais forte rede de conexões com empresas em 2010 no que diz respeito ao número de doações financeiras de pessoas jurídicas realizadas ao diretório nacional. Ele é seguido pelo PSDB e depois pelo PV. Apesar do pequeno número de nós, 28, o PV aparece em posição central na rede porque seus doadores também doaram para outros partidos, ou seja, as empresas mais centrais (transitivas) “puxaram” o PV para o centro da rede. Ao contrário do que aconteceu com PMDB e mais notadamente com PSDC e PRTB, cujos doadores não são centrais na rede.

As estatísticas individuais dos *clusters* mostram que a partir do G3, os números de vértices, distância geodésica média e densidade não são os mesmos nos dois modelos. As diferenças de resultados entre os dois métodos acontecem nos três *clusters* iniciais. A densidade da rede do PT é a menor e, proporcionalmente ao número de vértices, muito próxima da densidade do PSDB. Já o PV apresenta uma densidade significativamente maior. No entanto, a distância geodésica média de cada um deles é muito próxima entre si no primeiro modelo. Já no segundo, eles fazem parte de um mesmo *cluster*, não sendo possível identificar as características individuais de cada um. Como se vê, não basta plotar um desenho de rede, com distinção por cores de atributos ou interpretar os coeficientes isoladamente. É preciso saber o que cada informação representa de fato. No primeiro *cluster*, a resposta dá-se em função da pergunta: quais partidos recebem doações de quais empresas? Independentemente de serem doações a apenas um partido ou a mais de um. Aliás, a formação dos *clusters* indicará se dois ou mais partidos apresentam o mesmo tipo de doador. No segundo modelo, a pergunta é: como os doadores se aproximam ou se distanciam de outros doadores em relação aos partidos? Aqui, os partidos tendem a fazer parte de *clusters* distintos. A não ser que

o grau de vizinhança entre eles seja muito alto. Nesse caso, teremos como resposta a identificação de partidos que são mais parecidos do que diferentes no que diz respeito à proximidade com empresas que doam recursos para campanhas eleitorais.

5.5 Referências bibliográficas do Capítulo V

Bourdieu, P. (2003). Introdução a uma sociologia reflexiva. In: *O poder simbólico* (6ª Edição). Rio de Janeiro: Bertrand Brasil.

Burt, R. S. (1984). Network Itens and the General Social *Survey*. *Social Networks*, 6, 293-339.

Costa, J. H. (2011). Entre a estrutura e a ação, melhor a relação: para pensar a análise de redes sociais. *Revista Espaço Acadêmico*, 10(117), 123-131.

Freeman, L. C. (1979). Centrality in Social Networks: Conceptual clarification. *Social Networks*, 1, 215-239.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of California.

Marques, E. C. L. *et al.* (2007). Dossiê: métodos e explicações da política. *Revista Brasileira de Ciências Sociais*, 22(64), 140-145.

Smith, M. A. *et al.* (2009). Analyzing (Social Media) Networks with NodeXL. Paper. *ACM, C&T´09*. University Pensylvania, USA. Disponível em: http://www.cs.umd.edu/hcil/trs/2009-11/2009-11.pdf. Acesso em fevereiro de 2018.

5.6 Exercícios propostos do Capítulo V

5.6.1 A partir do banco de dados BDCAP5V2_COLIG, disponível em https://blogempublico.files.wordpress.com/2018/02/bdcap5v2_colig.xlsx (lembre-se de que é necessário ter instalado o NodeXL antes de abrir o arquivo), considere as seguintes características:

O banco de dados possui as seguintes variáveis: 1) partido de todos os candidatos a prefeito em 2016 e 2) escolaridade dos candidatos a prefeito. Essa segunda variável foi categorizada em: i - até ensino fundamental; ii - ensino médio; iii - ensino superior. Trata-se de uma rede não direcional, pois consideramos que os partidos não definem seus candidatos por escolaridade e os candidatos não se filiam aos partidos em função do nível escolar. Como atributos, o banco de dados oferece as seguintes informações complementares: a) Unidade da Federação, b) Região do País, c) Tamanho do Município em categorias nos quais os candidatos disputaram a eleição municipal e d) resultado da eleição, se o candidato foi eleito ou derrotado. A partir dessas informações, faça o que segue:

A) Gere um gráfico de redes usando o algoritmo *Harel-Koren* para todos os casos e em seguida gere um *cluster* pelo algoritmo *Clauset-Newman-Moore* para analisar as principais estatísticas obtidas (distância geodésica média e densidade).

B) Gere um gráfico apenas para os eleitos, gere um novo *cluster* por vizinhança e analise as estatísticas.

C) Explique a diferença no número de *clusters* entra A e B, considerando as variáveis do teste.

Capítulo VI

Testes de Correlação

A associação entre diferentes características individuais é o primeiro passo para a análise bivariada e, muitas vezes, o único necessário.

Até aqui foram discutidas técnicas de análise predominantemente aplicadas a dados categóricos, para experimentos entre variáveis, entre casos ou entre casos e variáveis. Este capítulo dá início à apresentação de uma série de testes que são recomendados para quando se está trabalhando com variáveis intervalares ou contínuas. Aqui, começamos com os testes de correlação para, nos capítulos seguintes, passarmos às análises de regressão e seus desdobramentos em técnicas específicas para determinados objetos de pesquisa. São os chamados testes paramétricos, pois eles precisam de pelo menos uma variável intervalar para extrair medidas de tendência central e variabilidade com o fim de parametrizar as análises. Também veremos algumas adaptações estatísticas das análises paramétricas, com variáveis intervalares ou contínuas, para variáveis categóricas. Todas as análises deste capítulo foram realizadas no *RCommander* ou *RStudio*.

6.1 Correlação Linear Simples

Podemos entender uma correlação como um procedimento estatístico usado

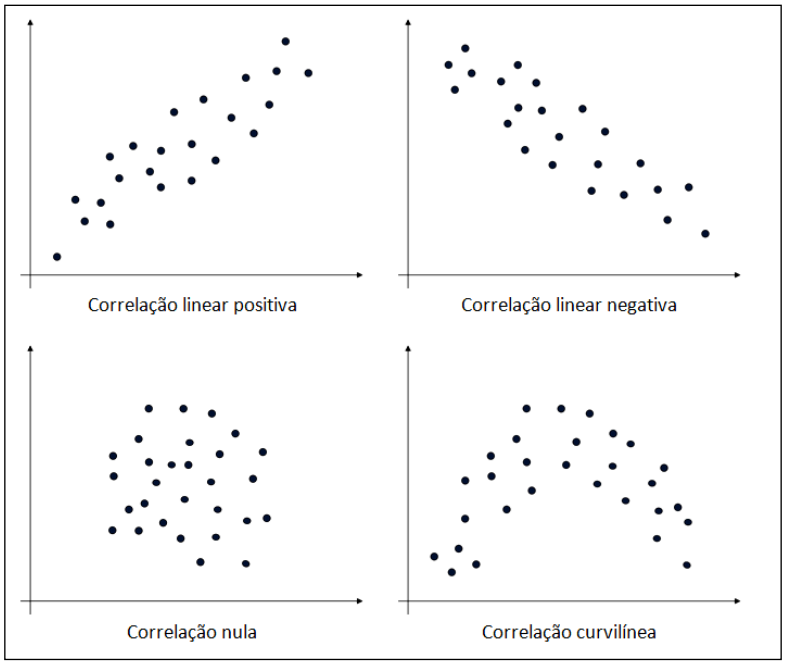
para medir e descrever a relação entre duas variáveis, onde as variações de uma podem ser usadas para explicar (antecipar) os valores da outra. Se há correlação entre duas variáveis, significa que elas podem ser usadas em um modelo de regressão, então, correlação é o primeiro passo para verificação da existência de relações consistentes entre duas variáveis. Uma correlação entre duas variáveis deve ser interpretada como a relação que existe entre elas, ou então, em outras palavras, a variação concomitante para as duas. O coeficiente de correlação é um coeficiente padronizado, pois seus valores variam entre -1 e +1, sendo que os extremos indicam correlações fortes, positivas ou negativas, e o zero indica a inexistência de correlação entre as variáveis. Quando duas variáveis estão correlacionadas podemos usar a variação de uma para predizer a variação da outra (Bunchaft & Kellner, 1999). Em outras palavras, um valor da variável X pode ser usado para predizer um valor da variável Y, se elas estiverem correlacionadas. Quando plotados em um gráfico de dispersão, os valores das variáveis permitem a produção de uma reta que vai indicar quais são os valores teóricos da distribuição conjunta das duas variáveis, caso a correlação seja perfeita.

A correlação mais conhecida é a correlação linear, ou correlação produto-momento de Pearson. Ao contrário dos testes apresentados nos dois primeiros capítulos deste volume, cujas medidas eram propostas para correlações entre variáveis categóricas, a correlação linear de Pearson é uma medida de inter-relação entre pelo menos duas variáveis contínuas. Sendo assim, podemos usá-la para saber se existe relação entre velocidade e compreensão leitora de estudantes brasileiros ou entre volume de leitura e rendimento acadêmico, por exemplo. A partir da correlação produto de Pearson, foram produzidos testes específicos para outros tipos de variáveis. Neste capítulo, além da correlação de Pearson, retomaremos alguns testes já apresentados para variáveis categóricas e apresentaremos outros mais usados para relações entre variáveis nominais ou ordinais.

Uma análise de correlação tem o objetivo de estimar numericamente o grau de relação entre os comportamentos de duas características de indivíduos de uma população ou de uma amostra da população, caso o objetivo seja fazer inferências estatísticas. Para fazer a análise da relação entre duas variáveis, dependemos do ponto de interseção entre duas variáveis em um espaço bidimensional (Triola, 1999). Esse par de pontuações pode ser representado sobre um eixo de coordenadas, chamado de

diagrama de dispersão. O diagrama ajuda a visualizar as relações entre as variáveis e mostra, graficamente, se a relação entre elas é linear ou curvilínea, o que determinará a viabilidade do uso da correlação linear de Pearson, como demonstram os exemplos abaixo, na figura 6.1:

Figura 6.1. Exemplo de tipos de distribuições teóricas



Fonte: autor

A partir do grau de relação entre duas variáveis apontado no teste de correlação linear, pode-se identificar se é adequada a postulação lógica sobre a existência de relação entre duas ou mais variáveis. Mas, atenção, a existência de correlação entre duas variáveis não implica em causalidade de uma sobre a outra. Mostra apenas que elas estão associadas. Ou seja, quando uma varia, isso tende a acontecer com a variação de outra ao mesmo tempo. O valor do coeficiente, que indica a magnitude da correlação, depende de vários fatores: se a amostra é aleatória ou representativa. Amostras não aleatórias, por exemplo, tendem a gerar coeficientes de correlação menores, pois a variação dos valores em uma das variáveis será menor. Por exemplo: correlacionar capacidade de memorização com nível de inteligência. Se a amostra tiver apenas os mais inteligentes, a correlação será menor. Outra característica tem a ver com o nível de significância dos resultados. Como o teste mede a associação conjunta das variações,

quando maior o número de casos, maior a chance de o resultado ser estatisticamente significativo. Por isso é preciso tomar cuidado ao se verificar o grau de significância de um teste de correlação de grandes conjuntos de dados. Por fim, é preciso considerar que o coeficiente de correlação é apenas uma estatística de sumarização, equivalente à média, por exemplo. Ele é incapaz de representar a dinâmica de todos os indivíduos. Existem diferentes tipos de testes de correlação, dependendo do tipo de variável utilizada, como indicado no quadro 6.1.

Quadro 6.1. Tipos de testes por tipo de variáveis a serem testadas

• Para duas variáveis contínuas	= Correlação produto-momento de Pearson;
• Para uma contínua e outra binária	= Teste de Correlação bisserial;
• Para duas variáveis binárias	= Coeficiente Phi e Q de Yule
• Para duas variáveis nominais	= Teste de correlação de Spearman;
• Para duas variáveis ordinais	= Teste de contingência Gama e T de Kendall.

Fonte: adaptado de Bunchaft e Kellner, 1999.

Para aplicar o teste produto-momento de Pearson a duas variáveis contínuas, o primeiro passo é identificar se existem ou não relação espacial através de um diagrama de dispersão, que representa visualmente uma relação entre duas variáveis. A partir de então a mesma relação pode ser sumarizada em um coeficiente de correlação. O coeficiente mais conhecido para variáveis contínuas é o de Pearson, representado pela letra "r". Se uma ou mais variáveis do teste forem categóricas nominais, deve-se usar o coeficiente de *Spearman*. Já se as duas variáveis forem categóricas ordinais, o coeficiente recomendado é aquele obtido no teste de Contingência. O objetivo de todo coeficiente de correlação é identificar a intensidade ou grau da correlação entra as variáveis testadas.

O coeficiente "r" pode ser calculado a partir de duas fórmulas: a fórmula *raw* ou a fórmula de *z-score*. Um coeficiente de correlação nada mais é do que a soma dos produtos cruzados e a covariância dos valores. Ou seja, "r" indica o grau com que X e Y variam juntas, relativo ao grau com que X e Y variam de maneira independente. Em termos matemáticos, seria:

$$r = \frac{(covariância\ X\ e\ Y)}{(variância\ de\ X\ e\ de\ Y)}$$

Duas variáveis são consideradas independentes quando não existe relação entre elas, ou seja, quando a correlação é nula. Se apresentar alguma correlação estatisticamente significativa, diz-se que existe uma dependência das variações delas, que pode ser positiva ou negativa, conforme indicação do sinal do coeficiente de correlação. Mas, atenção, podem-se encontrar correlações significativas que não representam autêntica dependência. São as chamadas correlações espúrias (Vasconcellos, 2000), como é, por exemplo, a relação entre número de ninhos de cegonha e número de nascimento na Islândia em determinados períodos do ano.

Em termos matemáticos, uma das fórmulas para obtenção do coeficiente de correlação é a seguinte. Ela utiliza apenas os valores originais e dispensa a inclusão das variações no cálculo. Existem outras fórmulas que podem gerar coeficientes um pouco mais precisos, porém, para os fins deste manual, a fórmula que segue é suficiente:

$$r = \frac{n.(\sum x.y) - (\sum x . \sum y)}{\sqrt{[n.\sum x^2 - (\sum x)^2].[n.\sum y^2 - (\sum y)^2]}}$$

Vejamos a aplicação da fórmula de correlação para o seguinte caso: usando o banco de dados das eleições de 2016 para as capitais brasileiras, identificaremos a correlação de Pearson entre o percentual de abstenção e o percentual de votos nulos para prefeito no primeiro turno. Busca-se identificar a existência de correlação entre a percentagem dos que decidem não participar da votação e a percentagem dos que participam, mas votam nulo. Se houver correlação e ela for negativa, significa que em capitais com maior percentual de abstenção há menos votos nulos, indicando que quem não tinha candidato decidiu não votar. Se a correlação for positiva, então, quanto maior a abstenção, maior tende a ser o voto nulo, indicando que esse tipo de voto é uma manifestação do eleitor que quer participar, mas não vê entre as alternativas de candidatos um que mereça seu voto. As duas variáveis são contínuas. Para calcularmos o coeficiente de correlação é preciso ter as somatórias dos valores de cada variável e seus valores elevados ao quadrado. A tabela 6.1 a seguir traz os dados originais acrescidos das colunas de x^2, y^2 e x.y.

Tabela 6.1. Abstenção e voto nulo em capitais para 1º turno da eleição para prefeito em 2016

Capital	%Abst. (X)	%V.Nulo (Y)	X^2	Y^2	X*Y
PORTO VELHO	18,64	15,37	347,28	236,32	286,48
RIO BRANCO	15,89	4,7	252,42	22,09	74,68
MANAUS	8,59	6,04	73,812	36,44	51,86
BOA VISTA	16,78	7,97	281,57	63,58	133,8
BELÉM	19	5,27	361,09	27,81	100,22
MACAPÁ	16,34	5,68	266,89	32,28	92,82
SÃO LUÍS	14,07	4,49	197,85	20,14	63,13
TERESINA	11,73	6,37	137,57	40,63	74,76
FORTALEZA	17,04	5,86	290,22	34,38	99,89
NATAL	19,6	12,56	384,26	157,79	246,24
JOÃO PESSOA	11,37	10,06	129,16	101,22	114,34
RECIFE	11,31	7,78	127,8	60,57	87,98
MACEIÓ	17,08	8,65	291,88	74,88	147,84
ARACAJU	18,04	16,6	325,31	275,71	299,49
SALVADOR	21,25	10,3	451,49	106	218,94
BELO HORIZONTE	21,66	14,28	469,26	203,94	309,36
VITÓRIA	10,76	6,79	115,77	46,041	73
RIO DE JANEIRO	24,28	12,76	589,46	162,86	309,84
SÃO PAULO	21,84	11,35	476,84	128,83	247,85
CURITIBA	16,44	9	270,29	80,91	147,88
FLORIANÓPOLIS	12,25	7,12	149,94	50,72	87,215
PORTO ALEGRE	22,51	8,88	506,55	78,86	199,86
CAMPO GRANDE	19,2	7,49	368,72	56,09	143,81
CUIABÁ	19,91	9,36	396,54	87,53	186,31
GOIÂNIA	20,83	7,11	434,02	50,49	148,03
PALMAS	15,58	7,14	242,84	50,95	111,23
SOMA	441,97	228,98	7.938,90	2.287,35	4.056,98

Fonte: autor a partir de TSE.

Com os dados tabulados, basta aplicar a fórmula apresentada acima:

$$r = \frac{n.(\sum x.y) - (\sum x.\sum y)}{\sqrt{[n.\sum x^2 - (\sum x)^2].[n.\sum y^2 - (\sum y)^2]}}$$

$$= \frac{26\ x\ 4.056{,}99 - (441{,}97\ x\ 228{,}98)}{\sqrt{[26\ x\ 7.938{,}95 - (441{,}97)^2]x[26\ x2287{,}35 - (228{,}98)^2]}}$$

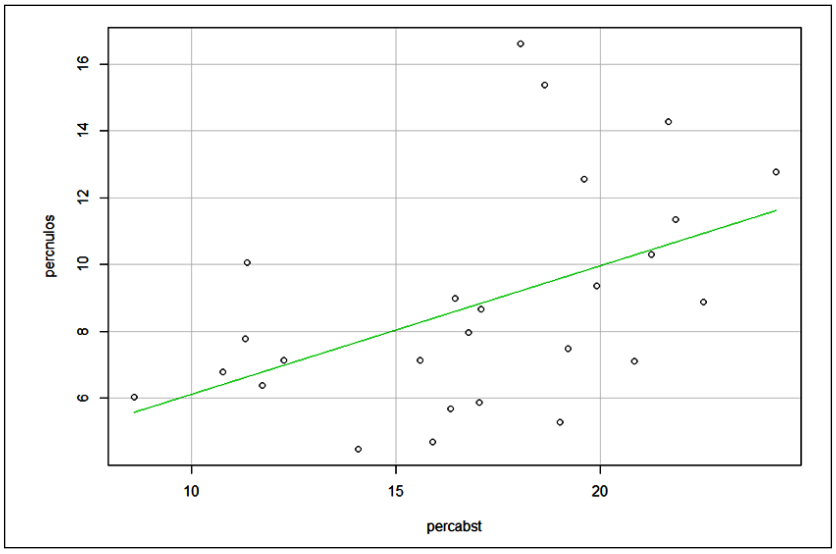
$$= \frac{105.481{,}74 - 101.211{,}13}{\sqrt{[206.412{,}7 - 195.337{,}4]\ x\ [59.471{,}1 - 52.441]}}$$

$$= \frac{4.270{,}61}{\sqrt{11.075{,}3\ x\ 7.031{,}1}} = \frac{4.270{,}61}{\sqrt{11.075{,}3\ x\ 7.031{,}1}} = \frac{4.270{,}61}{8.824{,}4} = \mathbf{0,483}$$

$$\cong \mathbf{48,3\%}$$

O coeficiente de correlação de Pearson de 0,483, que representa 48,3% de associação das variações, indica que há uma variação positiva entre as duas variáveis. Em municípios onde cresce a abstenção também aumenta os votos nulos e a proporção da associação é de 48,3%. Outra forma de representar a correlação é via gráfico de dispersão, como no gráfico 6.1 a seguir. Cada ponto representa uma capital. O eixo X indica o percentual de abstenção e o eixo Y o percentual de votos nulos no primeiro turno da eleição para prefeito de 2016.

Gráfico 6.1. Dispersão entre abstenção e voto nulo nas capitais brasileiras em 2016



Fonte: autor a partir de dados do TSE.

A reta de regressão traçada no gráfico de dispersão acima mostra que a direção da relação é positiva. Ou seja, conforme aumenta o percentual de abstenção tende a crescer o percentual de votos nulos no município. Um requisito importante para a utilização do coeficiente de correlação de Pearson é que a relação entre as variáveis X e Y seja linear. Além disso, é preciso que a medição seja pelo menos no nível intervalar e que no caso de inferências de uma amostra para a população, que exista aleatoriedade no sorteio da amostra. Podemos perceber pelo gráfico acima que existe uma distribuição desigual entre os casos que estão acima e abaixo da reta. Os casos que estão abaixo da reta apresentam menor dispersão, pois suas distâncias em relação à reta são menores que as distâncias dos casos que se localizam acima dela. Isso é um limitador para a interpretação dos resultados do teste.

Outra maneira para se calcular o coeficiente “r” é através da fórmula *Z-score*. Neste caso é preciso conhecer o Desvio Padrão em relação à média para encontrar o valor de Z nas duas variáveis.

$$r = \frac{\sum(Z_{x\ X}\ Z_y)}{N}$$

Onde:

Zx = (X-μ) / Dp

ZY = (Y-μ) / Dp

Para encontrar o valor de Z de cada unidade é preciso subtrair a unidade pela média e dividir pelo desvio padrão. No nosso exemplo, a variável Percentual de Abstenção apresenta μ= 19,99 e Desvio Padrão= 4,12 e a variável Percentual de Votos Nulos tem μ= 8,80 e Desvio Padrão= 3,28. Aplicando a fórmula temos que:

$$r = \frac{\sum(Z_{x\ X}\ Z_y)}{N} = \frac{12{,}10}{26} = 0{,}465\ \cong 46{,}5$$

É possível perceber que houve uma diferença nos coeficientes calculados pelas duas fórmulas, passando de 48,3% na primeira, para 46,5% agora. Essa diferença do coeficiente “r” da fórmula de *Z-score* para a fórmula anterior, de 1,8 ponto percentual, deve-se a dois fatores. O primeiro é o arredondamento de valores durante os cálculos. O segundo é que pela segunda fórmula levamos em conta uma medida de variação, o desvio padrão, e isso aumenta a precisão do coeficiente, principalmente quando se está trabalhando com distribuições que apresentam grandes variações internas.

Todo coeficiente de correlação será um valor compreendido entre -1,00 e +1,00. À medida que se aproxima de (±)1,00, a intensidade da correlação aumenta. O sentido da correlação é expresso pelo sinal que apresenta seu coeficiente, sendo negativo para correlações em direção oposta e positiva para correlação na mesma direção. Existem diferentes formas de interpretação desses valores, a mais comum é a seguinte:

Quadro 6.2. Intervalos para interpretação da força da correlação

COEFICIENTE	INTERPRETAÇÃO
r = (±) 1,00	Correlação perfeita
(±) 0,80 < r < (±) 1,00	Muito alta
(±) 0,60 < r < (±) 0,80	Ata
(±) 0,40 < r < (±) 0,60	Moderada
(±) 0,20 < r < (±) 0,40	Baixa
0,00 < r < (±) 0,20	Muito baixa
r = 0	Nula

Para Gulford (1965), em relação a estudos teóricos, qualquer correlação, ainda que pequena, desde que seja estatisticamente significativa, indica um grau de relação entre as duas variáveis. No exemplo da relação entre abstenção e votos nulos em capitais brasileiras, o coeficiente de 0,483 é considerado moderado.

Por fim, em qualquer um dos casos a recomendação é que ao se trabalhar com estatística descritiva, ou seja, testar a correlação usando dados de toda a população, que a divisão seja feita por (N), mas, quando a correlação for entre elementos amostrais, ou seja, para servir como estatística inferencial, a divisão deve ser feita por (N-1). Além disso, é preciso respeitar alguns pressupostos antes de analisar os coeficientes de uma correlação:

1 – As duas variáveis correlacionadas devem apresentar distribuições normais. Para detectar quebra de normalidade basta gerar um histograma, gráfico QQ (discutido mais adiante) ou verificar as estatísticas sumarizadoras;

2 – A relação entre as duas variáveis deve ser linear e não em outra forma. Para detectar quebra de linearidade da relação basta gerar um gráfico de dispersão e analisá-lo;

3 – As distribuições dos erros em relação à reta devem ser homoscedásticas. A homoscedasticidade também pode ser verificada no gráfico de dispersão. Trata-se da homogeneidade nas distâncias verticais entre os pontos e a reta. Essa distância é chamada de resíduo da correlação e os resíduos precisam estar uniformemente distribuídos ao longo de toda a reta para que o pressuposto da homoscedasticidade não seja quebrado.

4 – Também é preciso que exista confiabilidade (*reliability*) nas relações entre X e Y para que um coeficiente de correlação seja considerado válido. A confiabilidade é quebrada por erro ou viés, que é uma tendência de desvio consistente na mesma direção. Por exemplo, podemos chamar de $\bar{Y}$ o valor de Y que está exatamente sobre a reta da

correlação. Na maioria das vezes $\bar{Y}$ é apenas um valor teórico, pois na realidade Y está fora da reta. No mundo ideal, $\bar{Y}$ deveria ser igual a Y. No entanto, o valor de $\bar{Y}$ é composto por: $\bar{Y}$ =valor real de Y + viés + erro. Quanto mais distante o valor de Y for do valor real, maior será o viés e o erro, portanto, menos confiável será a medida. Existem testes para verificar ocorrência de viés e medir a confiabilidade da correlação entre as variáveis.

5 – É preciso existir validade na relação entre X e Y. Nem sempre um coeficiente de correlação alto indica uma relação válida entre X e Y. Muitas variáveis correlacionadas não são naturais, mas constructos humanos. Um constructo é um objeto ideal, que não pode ser observado diretamente, o oposto de objetos reais. Por exemplo, "nível de satisfação com o governo" é um constructo usado em muitas correlações e que não pode ser materializado. Então, precisamos tomar alguns cuidados para tornar um constructo observável e quantificável.

Para operacionalizar um constructo é preciso verificar quatro tipos de validades nesse constructo:

a) validade de conteúdo: o conteúdo desse constructo deve ser de conhecimento geral. A população deve entender de forma parecida o que significar o termo "satisfação com governo", por exemplo;

b) validade de convergência: quando a variação do constructo é correlacionada com outro similar. Por exemplo, satisfação com governo correlacionado com simpatia pelo partido do governante;

c) validade de divergência: quando a variação do constructo não está correlacionada com outro não similar a ele. Por exemplo, satisfação com governo correlacionado com simpatia por determinado time de futebol;

d) validade monológica: quando as variações do constructo estão em acordo com o que a literatura ou a teoria da área vem apresentando a respeito do assunto. Por exemplo, a correlação entre satisfação com o governo e os critérios de escolha racional para avaliação de governos que são usados quando há algum benefício individual identificado nesse governo (Bunchaft & Kellner, 1999). Para verificar se os pressupostos indicados acima estão mantidos no teste recomenda-se: i) gerar um histograma para verificar a forma da distribuição dos casos; ii) gerar as estatísticas sumárias e analisar seus resultados.

6.2 Aplicação dos testes de correlação para amostras

Normalmente, a correlação de Pearson é aplicada para se encontrar relações concretas em uma população, porém, pode ser usada para a estimação das relações a partir de uma amostra. Nesse caso é preciso que as amostras sejam representativas e aleatórias, o que acontece quando todas as distribuições são equivalentes e perfeitamente iguais em todas as partes ou características medidas. A interpretação dos coeficientes ajuda a identificar intuitivamente a relação entre duas variáveis, porém, para a análise estatística é preciso usar o grau de significância, que está em função do coeficiente obtido e do número de indivíduos. Esses índices de significância indicam os valores a partir dos quais se deve rejeitar a hipótese nula. Um coeficiente pode apresentar determinada correlação, porém, apenas se esse valor estiver abaixo do nível de significância de 0,050 poderemos dizer que a correlação é válida para extrapolações de amostras à população, pois nesse caso nada se opõe a rejeitar a hipótese nula. Como muitas vezes estamos trabalhando com amostras e não população, além de conhecer o coeficiente de correlação, é importante saber se aquele resultado pode ser extrapolado para um grupo maior de indivíduos caso seja usado em um procedimento inferencial. Ou seja, se o coeficiente obtido tem significância estatística para extrapolar. Para isso, são usados como parâmetros os valores críticos de significância aplicados aos coeficientes, como constam na tabela 6.2 abaixo.

Tabela 6.2. Limites críticos de significância estatística do Coeficiente de Correlação de Pearson

Valores Críticos do Coeficiente de Correlação de Pearson (r)								
N	**α = 0,05**	**α = 0,01**	**N**	**α = 0,05**	**α = 0,01**	**N**	**α = 0,05**	**α = 0,01**
4	0,95	0,999	**13**	0,553	0,684	**30**	0,361	0,463
5	0,878	0,959	**14**	0,532	0,661	**35**	0,335	0,430
6	0,811	0,917	**15**	0,514	0,641	**40**	0,312	0,402
7	0,754	0,875	**16**	0,497	0,623	**50**	0,279	0,361
8	0,707	0,834	**17**	0,482	0,606	**60**	0,254	0,330
9	0,666	0,798	**18**	0,468	0,590	**70**	0,236	0,305
10	0,632	0,765	**19**	0,456	0,575	**80**	0,220	0,286
11	0,602	0,735	**20**	0,444	0,561	**90**	0,207	0,269
12	0,576	0,708	**25**	0,396	0,505	**100**	0,196	0,256

Fonte: Bunchaft e Kellner, 1999.

A tabela 6.2 indica o valor acima do qual o coeficiente de correlação deve se encontrar para ser considerado estatisticamente significativo. Para tanto, leva em consideração o número de casos do cálculo (N) e o nível de significância (α), que pode ser de 0,050 para intervalo de confiança de 95% e 0,010 para intervalo de confiança de 99%. Aplicando ao nosso exemplo, com r = 0,483 na primeira fórmula e N = 26, precisamos identificar o limite crítico do coeficiente para o (N) mais próximo do nosso. Encontramos para n = 25 um limite crítico de 0,396 para α de 0,050. Portanto, o nosso coeficiente está acima do limite crítico. Se estivéssemos trabalhando com uma amostra aleatória de casos, poderíamos extrapolar o resultado para toda a população.

É possível medir o erro de amostra na correlação para verificar a representatividade dessa amostra. Os erros de amostras são definidos pelas diferenças ou distâncias entre os valores amostrados e os valores da população. O problema é que quando estamos trabalhando com estatística inferencial raramente conhecemos os parâmetros para a população. Então, como fazer para conhecer o erro de amostra? Existem alguns fatores que ajudam a determinar o erro amostral. O primeiro é o tamanho da amostra: quanto maior o tamanho da amostra em relação ao tamanho da população, menor será o erro amostral. O segundo é a variância ou a heterogeneidade da característica medida na correlação. Quanto maior a variância na população, maior a possibilidade de erro amostral. Portanto, o erro de amostra é estimado pelo tamanho da amostra e pela variância da característica medida na população.

Uma das formas para se medir o erro amostral em uma correlação é através do coeficiente de **erro-padrão do r**, que permite estabelecer o intervalo de confiança de r, ou seja, entre que valores o coeficiente poderá oscilar na população para um determinado nível de significância. A fórmula é:

$$SE = \frac{DP}{\sqrt{N}}$$

Onde:

SE = erro padrão

DP = Desvio Padrão

N = tamanho da amostra

Como o resultado estima a quantidade média de erro da amostra, ele é usado para identificar os limites (superior e inferior) da faixa em que o coeficiente de correlação se encontrará, caso o teste fosse aplicado a toda população. Aplicando ao nosso exemplo, para a variável percentual de abstenção nas capitais em 2016, teríamos:

$$SE_{abs} = \frac{SD}{\sqrt{N}} = \frac{4{,}12}{\sqrt{26}} = 0{,}809$$

Se estivermos usando o intervalo de confiança de 95%, o valor padronizado para o cálculo será de 1,96. Assim, o cálculo para o limite superior e inferior do intervalo de confiança seria de:

Limite inferior: Média – (1,96 x SE) = 16,99 - (1,96 x 0,809) = 12,15
Limite superior: Média + (1,96 x SE) = 16,99 + (1,96 x 0,809) = 15,33

O resultado é que se fosse uma amostra de 26 casos, com erro padrão de 0,809, intervalo de confiança de 95% e média de 16,99, o percentual de abstenção para a população de municípios estaria entre o mínimo de 12,15% e o máximo de 15,33%. No próximo tópico, discutiremos outro conjunto de testes, os de determinação e alienação.

6.3 Coeficiente Linear de Determinação e de Alienação

O passo seguinte à identificação da correlação entre duas variáveis é verificar quanto que há de determinação da variação de uma sobre a outra. Uma correlação nos oferece as associações entre variações de variáveis e **nunca** deve ser interpretado como uma determinação. O coeficiente de determinação indica quanto da variação de Y em relação à média de variação de X é explicada pelo modelo construído. Nesse caso, o modelo deve ser linear. Ele é interpretado como a proporção da variabilidade de uma variável Y (dependente) que pode ser explicada pela variabilidade da variável X (independente). A partir do coeficiente de determinação já podemos falar sem problemas entre variáveis dependentes e independentes, pois a partir daqui estamos buscando ex-

plicações sobre quando a variação de uma variável determina de variação em outra, ou seja, quanto que a segunda depende da primeira (Triola, 1999). Obtém-se o coeficiente de determinação elevando o coeficiente de correlação r ao quadrado. Portanto:

$$\text{Coeficiente de Determinação} = r^2$$

Se multiplicarmos o r^2 por 100, o coeficiente de determinação pode ser interpretado como a porcentagem da variabilidade conjunta entre as duas variáveis. No caso do exemplo anterior, da relação abstenção e voto nulo, o coeficiente de determinação poderia ser calculado a partir da correlação, da seguinte forma:

$$\text{Coeficiente de Determinação: } r^2 = 0{,}483^2 = 0{,}233 = 0{,}233 \times 100 \cong 23{,}3\%$$

Portanto, podemos dizer que 23,3% da variação dos votos nulos para prefeito nas capitais brasileiras é explicada pela variação de abstenção em 2016, de acordo com o modelo apresentado. Como a correlação sempre é uma fração entre zero e 1,00, ao elevá-la ao quadrado, o resultado da determinação será um valor menor que o anterior. Porém, só faz sentido falar em coeficiente de determinação caso exista uma causalidade lógica na relação entre as duas variáveis. Sem isso, deve-se falar em medida de associação (r) e não de determinação (r^2).

Já o coeficiente de alienação, representado aqui pela letra K, é a proporção da variabilidade de Y (variável dependente) que não é explicada pela variabilidade de X (variável independente). Sendo assim, para obter a alienação, basta subtrair de um o coeficiente de determinação:

$$\text{Coeficiente de Alienação: } K = 1 - r^2$$

Aplicando ao nosso exemplo, teríamos que:

$$\text{Coeficiente de Alienação: } k = 1 - 0{,}233 = 0{,}767 \times 100 \cong 76{,}7\%$$

A interpretação desse coeficiente é a seguinte: 76,7% da variação do percentual de votos nulos para prefeito nas capitais em 2016 não é explicada pela variação de abstenção naquela eleição. Se, além de identificar os coeficientes de correlação, determinação e alienação, que dizem respeito à dinâmica da relação entre X e Y, o pesquisador quiser fazer predições, ou seja, identificar variações possíveis a partir de valores já existentes, é necessário usar outro tipo de teste: a regressão. Trataremos dos testes de regressão no próximo capítulo. Mas, antes disso, precisamos apresentar os principais testes de quebra de pressupostos para testes lineares.

6.4 Pressupostos a serem respeitados em Análises de Correlação

Existem dois pressupostos que se repetem em todos os testes de regressão e, embora não seja prática comum, é fortemente recomendável a realização testes para verificar se as variáveis não estão quebrando os pressupostos antes de entrar nas regressões propriamente ditas. São eles:

a) existência de distribuição normal na variável dependente (Y); e

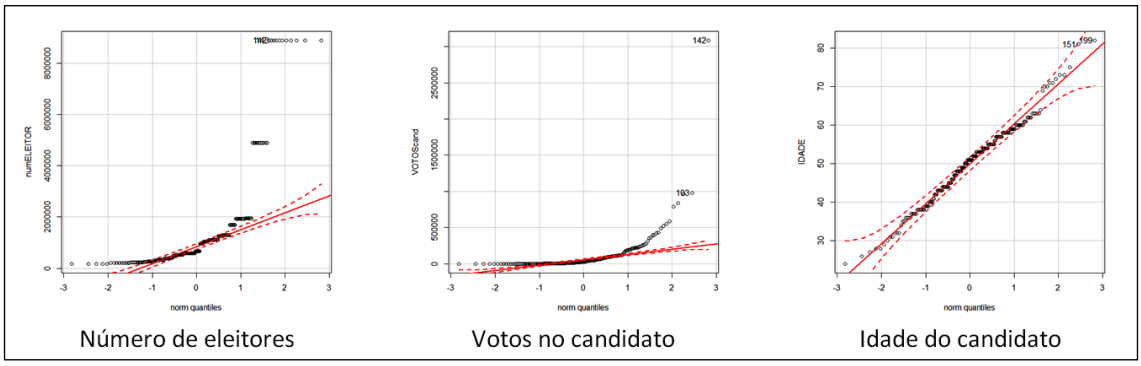
b) relação linear entre as variáveis preditoras e a variável a ser testada.

Trataremos da primeira aqui e da segunda no próximo capítulo. Para verificar se a distribuição de Y é normal, pode-se gerar um histograma ou analisar algumas estatísticas básicas. Para tanto, existe um gráfico específico que verifica esse tipo de distribuição que é o gráfico Q-Q (quantil-quantil). Bastante útil, é um teste de probabilidades que identifica a normalidade da distribuição da variável dependente. Nesse gráfico, os valores são distribuídos em relação a uma distribuição esperada, o que corresponderia à distribuição normal. Se todos os pontos do gráfico ficarem próximo à linha dos valores esperados, dentro do intervalo de confiança, significa que a distribuição é normal. Se forem identificados *outliers*, valores distantes da linha predita, o pressuposto da normalidade da distribuição dos casos da variável Y foi quebrado.

Um exemplo da verificação de normalidade de uma distribuição por gráfico Q-Q segue abaixo. São três gráficos com três variáveis distintas para ilustrar comportamentos possíveis em uma distribuição. As variáveis pertencem ao banco de dados de

características de todos os candidatos a prefeito de capitais do Brasil em 2016. Ao todo são 106 candidatos. Os gráficos foram gerados na função “Gráficos/Gráficos de Comparação de Quantis” da interface *RCommander* do *RStudio*. A primeira imagem é o gráfico QQ para o número de eleitores registrados nas capitais brasileiras em 2016. Perceba que ela não segue uma distribuição normal, pois a maior parte dos casos se distancia da reta e fica fora do intervalo de confiança indicado pelas linhas tracejadas. A segunda imagem é a distribuição da variável “votos obtidos no primeiro turno”. Perceba que ela também não tem uma distribuição normal, mas diferente da anterior, aqui existe um número pequeno de casos distantes da reta que indica a normalidade. São os candidatos de capitais com mais eleitores. O caso 142 está na cidade de São Paulo. Esse tipo de variável, com alguns casos “desgarrados”, não deve ser usado em regressões como dependente, pois quebra o pressuposto da normalidade de distribuição. Já o terceiro caso, a variável “idade do candidato”, percebe-se uma distribuição normal, com todos os casos distribuindo-se em torno da reta e ficando dentro do intervalo de confiança.

Gráfico 6.2. Exemplos de gráficos QQ para teste de normalidade da variável Y



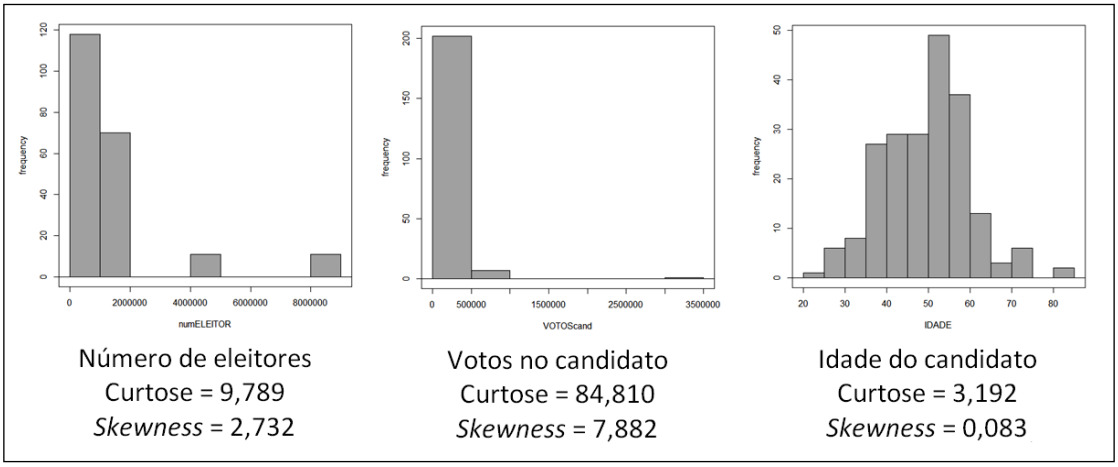
Fonte: autor a partir de TSE.

Nos dois primeiros casos, o número de *outliers* positivos, casos desviantes para cima, é grande, enquanto que para percentual de votos também há casos desviantes para baixo no meio da distribuição. Isso impediria a realização de testes estatísticos inferenciais a partir de uma amostra dessa distribuição. Por isso, os pontos observados praticamente assumem uma coluna vertical no início da distribuição, para só depois começar a se distribuir em torno dos valores esperados.

Outra forma de testar a normalidade da distribuição de uma variável é pelo his-

tograma. Se a opção for a análise das distribuições por histograma, a recomendação é olhar para a curtose, ou seja, o formato da distribuição (sobre formas de distribuição ver o volume I deste Manual). Uma distribuição é normal quando a curtose fica ao redor de 3,000. A medida de *Skewness* mostra a forma das caudas. Se *Skewness* se aproximar de zero, então as duas caldas serão equivalentes e a curva se aproximará da normal. Como o *RCommander* não calcula diretamente a curtose, rodamos os dados no *RStudio* e calculamos posteriormente para cada variável[1].

Gráfico 6.3. Histogramas, curtose e *Skewness* para teste de normalidade da variável Y



Fonte: autor a partir de TSE

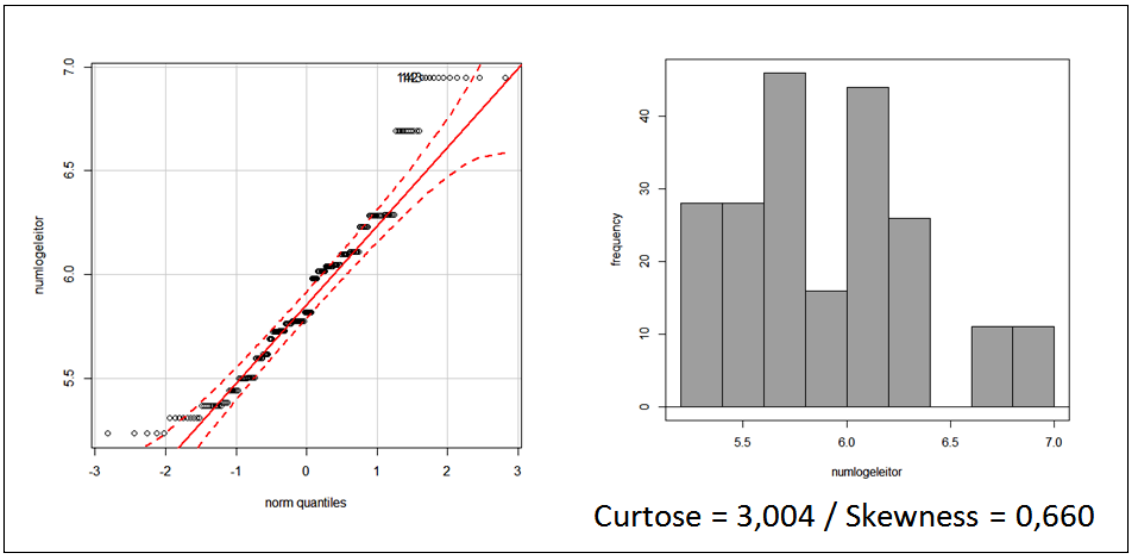
Nos casos em que há quebra do pressuposto da normalidade, não é possível utilizar a variável para a realização de testes estatísticos inferenciais (é óbvio que isso não se aplica quando se trata de dados para a população como um todo). Nos três exemplos acima, apenas a variável idade do candidato apresenta uma curva próxima à normal, com curtose de 3,192 e *Skewness* em 0,083. As outras duas variáveis apresentam curtose muito acima de 3,000 e *Skewness* bem diferente de zero. Nesses casos, é preciso usar alguma técnica de transformação de variáveis originais em variáveis que se aproximam da distribuição normal. São as chamadas técnicas de transformação de dados, que abordaremos a seguir.

[1] Para obter os valores de curtose e *Skewness* no *RStudio* é preciso instalar o pacote “moments”. Nele, as funções solicitadas são: Kurtosis (x) e *Skewness* (x), onde X é a variável analisada.

6.4.1 Transformações de dados para normalização de distribuições

Quando se constata a quebra do pressuposto da normalidade, é necessário fazer a transformação de dados para normalização de valores antes da realização de testes estatísticos com fins inferenciais. A transformação é uma função simples aplicada a todos os valores. A regra básica em uma transformação de dados para normalização é que a ordem dos casos na distribuição (rank) não pode ser alterada. Já a distância entre cada caso na distribuição, sim. As transformações mais comuns são: raiz quadrada, logaritmo, inverso. Apenas para dar um exemplo dos efeitos das transformações, o gráfico a seguir apresenta o gráfico QQ e o histograma para a variável "logaritmo do número de eleitores das capitais".

Gráfico 6.4. Gráfico QQ, histogramas, curtose e *Skewness* para log do número de eleitores

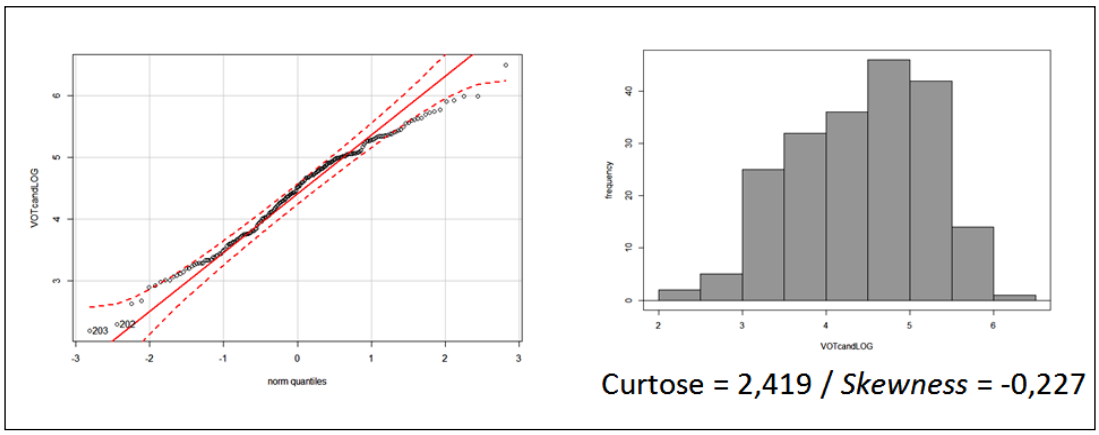


Como se pode perceber nas imagens acima, com a transformação logarítmica, os valores observados ficaram muito mais próximos da distribuição esperada normal, assim como a *Skewness* se aproximou de zero, a curtose ficou em 3,004. Portanto, após a transformação logarítmica da variável número de eleitores, ela poderia ser usada como variável dependente em um modelo de estatística inferencial sem quebrar o pressuposto da normalidade da distribuição dos resíduos da variável Y.

6.5 Aplicação da correlação de Pearson e outros coeficientes no *RCommander*

O banco de dados BDCAP7V2_CAPITAL[2] possui informações sobre o desempenho dos candidatos a prefeito em 2016 para as capitais brasileiras. Foram incluídos no banco todos os candidatos a prefeito que fizeram pelo menos um voto em uma das 26 capitais de estado brasileiras. Os dados estão em duas dimensões. Uma delas diz respeito às características dos candidatos e outra a dos municípios. Ao todo, são 203 candidatos no banco de dados. Para demonstrar a correlação de Pearson no *RCommander* usaremos as variáveis "idade do candidato" e "votação no 1º turno". O objetivo é testar se candidatos mais velhos tendem a ter mais votos que candidatos mais jovens. No tópico anterior já fizemos os testes de normalidade e constatamos que a variável idade não quebra o pressuposto da normalidade, já a votação não apresenta normalidade. Por isso, usaremos a transformação logarítmica da votação. Os gráficos QQ e histograma do log da votação seguem abaixo (gráf. 6.5.) e mostram que a variável transformada distribui-se na forma de uma curva normal, com curtose próxima a 3,000 e *Skewness* cerca de zero.

Gráfico 6.5. Gráfico QQ, histogramas, curtose e Skewness para log de votação dos candidatos

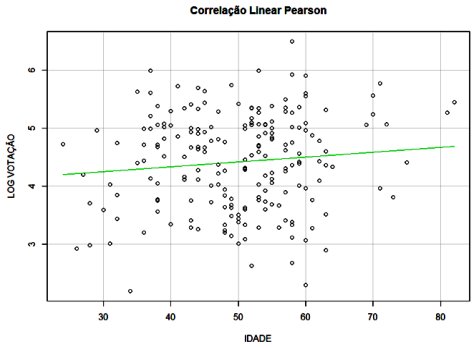


Com isso podemos fazer a correlação linear de Pearson entre idade e log da votação. No *RCommander* o caminho é: "Estatísticas/Resumo/Teste de Correlação". Na caixa, marcam-se as variáveis "idade" e "votcandlog" para correlação de Pearson.

[2] Disponível em: https://blogempublico.files.wordpress.com/2018/02/bdcap7v2_capital1.xlsx

A saída está no quadro abaixo. Os principais resultados da saída são os seguintes: a estatística "t" para a diferença de médias (1,525), graus de liberdade do teste (201) e o *p-valor* (0,128) que indica a significância estatística para a diferença de médias. Esse primeiro conjunto de dados nos mostra que a estatística "t" é muito baixa e se fosse uma amostra o *p-valor* não permitira a extrapolação de resultados, pois fica bem acima de 0,050. Em outras palavras, as variações de idade e log da votação são independentes entre si. A segunda linha de resultados mostra os limites do intervalo de confiança para 95%. No caso, o limite inferior fica em -0,031 e o superior em +0,241. Como o intervalo passa por zero, não temos segurança estatística suficiente para afirmar que as variações estão associadas na população, se estivéssemos trabalhando com uma amostra. Por fim, aparece o coeficiente de correlação de Pearson (0,107), que indica que para a população, 10,7% da variação da votação está correlacionada com a variação da idade dos candidatos. Trata-se de uma correlação muito baixa (ver tab. 6.2), porém, positiva, indicando que candidatos mais velhos tenderiam a ter mais votos.

```
Linha de Comando:
Rcmdr>  with(Capital, cor.test(IDADE, VOTcandLOG, alternative="two.
sided", method="pearson"))

Resultados:
      Pearson's product-moment correlation
data:  IDADE and VOTcandLOG
t = 1.5259, df = 201, p-value = 0.1286
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03116028 0.24116472
sample estimates:
      cor
0.1070087

Diagrama de Dispersão entre as duas variáveis:
```



Ao final da saída foi incluído o diagrama de dispersão com a reta dos mínimos quadrados para representar graficamente a associação entre as duas variáveis. O *RCommander* também oferece a possibilidade de mostrar os resultados na forma de matriz de correlação. Para isso o caminho é: "Estatísticas/Resumo/Matriz de Correlação". A saída está no quadro a seguir. A matriz de correlação pode ser formada por N variáveis, porém, sempre fornecerá os coeficientes de correlação entre pares. No nosso caso, o coeficiente de 0,107 aparece para as variáveis "idade" e "votcandlog". No *RCommander*, a matriz de correlação não traz as informações complementares do teste anterior, como o coeficiente do teste "t", *p-valor* e intervalo de confiança. Uma matriz de correlação serve para fazer explorações a partir da visualização rápida de correlações entre pares de variáveis.

```
Linha de comando:
Rcmdr>  cor(Capital[,c("IDADE","VOTcandLOG")], use="complete")

Resultados:
              IDADE VOTcandLOG
IDADE      1.0000000  0.1070087
VOTcandLOG 0.1070087  1.0000000
```

Como representado no quadro 6.1 deste capítulo, o tipo de variável determina o teste de correlação adequado como medida de associação. A correlação de Pearson se aplica apenas a quando temos duas variáveis contínuas. O *RCommander* oferece outros dois testes de correlação. O teste de *Spearman* para duas variáveis categóricas nominais e o teste Tau de *Kendall* para duas categóricas ordinais. Para exemplificar uma correlação de *Spearman* no nosso banco de dados vamos correlacionar "estado civil" do candidato, com as categorias: solteiro, casado, divorciado ou viúvo; e a variável "cor da pele" do candidato, com as categorias: branca, parda ou preta. O objetivo é ver se há alguma associação nos candidatos entre essas duas variáveis. O resultado está no quadro abaixo:

```
Linha de comando:
Rcmdr>  with(Capital, cor.test(COD_CPELE, COD_ESTCIV,
alternative="two.sided", method="spearman"))

Resultados: Spearman's rank correlation rho

data:  COD_CPELE and COD_ESTCIV
S = 1436300, p-value = 0.6687
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03021571
```

A estatística mais importante é o coeficiente de correlação de Spearman (*rho*), que ficou em -0,030. O valor é muito baixo, próximo a zero, portanto, correlação inexistente. Porém, se quiséssemos analisar a direção do sinal negativo, precisaríamos conhecer a categorização das duas variáveis. No caso, estado civil começa com solteiro e termina com viúvo. Cor da pele é categorizada na sequência: branca, parta, preta. Assim, como o sinal é negativo, a associação seria inversa, com tendência de termos mais candidatos com pele preta no grupo dos solteiros e mais cor da pele branca entre casados e viúvos.

Como exemplo para a correlação entre duas variáveis ordinais, T de *Kendall*, usaremos as variáveis “Escolaridade” do candidato, que começa no nível mais baixo de escolaridade (lê e escreve) e segue até o nível superior, e a variável “VotocandCat”, que é a organização das votações dos candidatos em quatro categorias ordinais (até 10%, de 10% a 30%, de 30% a 50% e acima de 50%). O objetivo do teste é verificar se candidatos com maior escolaridade tendem a estar nos grupos superiores e votação. Os resultados estão no quadro a seguir:

```
Linha de comando:
Rcmdr>  with(Capital, cor.test(COD_ESCOLAR, VOTCAND_CAT,
alternative=”two.sided”, method=”kendall”))

Resultados: Kendall’s rank correlation tau

data:  COD_ESCOLAR and VOTCAND_CAT
z = 3.2223, p-value = 0.001272
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.2102613
```

Aqui os resultados mostram uma correlação mais forte que a anterior. O coeficiente t = 0,210 ainda é considerado fraco, mas indica que 21% das variações de posição na votação estão associadas às variações de escolaridade. Como o coeficiente é positivo, as duas variáveis tendem a variar na mesma direção. Quanto mais escolaridade, maior a tendência de estar nos grupos superiores de votação. Além disso, o *p-valor* = 0,001 fica abaixo do limite crítico de 0,050 para intervalo de confiança de 95%, indicando que poderíamos extrapolar os resultados para a população, caso isso fosse uma amostra.

Um caso particular é quando se está usando variáveis de tipos diferentes. Por exemplo, correlacionar uma variável contínua com uma categórica nominal ou ordinal. Nesses casos, a recomendação é sempre usar o teste mais restritivo, pois um teste para variáveis categóricas é capaz de categorizar os valores contínuos em vizinhanças para fazer a correlação. Nesse caso, entre Pearson e Kendall, se uma das variáveis for contónua e outra categórica ordinal, prefira Kendall. Entre Kendall e Spearman, se uma for ordinal e outra nominal, prefira Kendall, pois ele estará considerando a transitividade das categorias, da menor para a maior, para as duas variáveis, embora seja necessário para apenas uma delas.

Por fim, na caixa dos testes de correlação do *RCommander*, oferece-se a possibilidade de marcar se a hipótese alternativa a ser testada é bilateral (vale tanto para valores positivos quanto para negativos), é menor que zero ou é maior que zero. O padrão é marcar teste bilateral, mas existem casos em que se quer verificar a correlação apenas para uma das caudas da distribuição. Nesse caso, utiliza-se a opção acima ou abaixo de zero para hipótese alternativa.

6.6 Referências bibliográficas do Capítulo VI

Bunchaft, G., & Kellner, S. R. O. (1999). *Estatística sem mistérios*. Volume II. Petrópolis: Editora Vozes.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

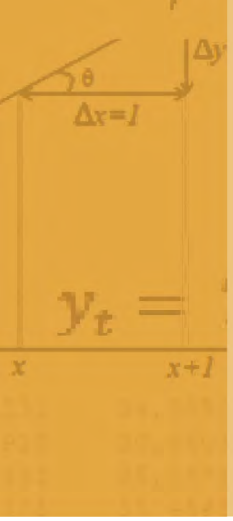
Triola, M. F. (1999). *Introdução à Estatística* (7ª edição). Rio de Janeiro: Editora LTC.

Vasconcellos, M. A. S., & Alves, D. (2000). *Manual de Econometria*. São Paulo: Editora Atlas.

6.7 Exercícios do propostos Capítulo VI

6.7.1 No banco de dados BDCAP7V2_CAPITAL, disponível em https://blogempublico.files.wordpress.com/2018/02/bdcap7v2_capital1.xlsx, faça os testes de normalidade da distribuição da variável “percentual de votos válidos” (numpervotval) gerando o QQ, o histograma, a medida de Curtose e de *Skewness*. Analise os resultados e responda se é possível usar essa variável diretamente em um teste de correlação. Se não, qual a alternativa recomendada?

6.7.2 Utilizando o mesmo banco de dados, considere as variáveis “votcand_cat” (votação obtida pelo candidato em categorias) e “idade” para rodar três testes de correlação: a) Pearson, b) Spearman, e c) Kendall. Responda: i) qual dos três está tecnicamento correto e por quê? ii) interprete as diferenças entre os coeficientes.



Capítulo VII

Princípios dos Testes de Regressão

A possibilidade de antecipar valores de uma variável a partir das variações de outra é um ganho na qualidade das pesquisas preditiva.

Uma vez conhecidos os princípios dos testes de correlação, que medem a associação entre variações de duas ou mais variáveis, agora vamos dar um passo adiante e, além da associação, identificar determinações. Ou seja, quanto que ao aumentar uma unidade na variável explicativa é possível predizer de aumento na variável dependente. Esse é o tema dos testes de regressão. Mas, antes de entrarmos nos testes propriamente ditos, vale a pena ocupar um tempo para entender o princípio que move as regressões. O método mais usado é o Método dos Mínimos Quadrados (MMQ) ou a sigla em inglês OLS[1]. Trata-se de uma técnica matemática que busca o melhor ajuste para um conjunto de dados qualquer, minimizando as diferenças entre unidades vizinhas pela soma dos quadrados da distância entre os valores observados e o valor estimado. Essas diferenças, ou distâncias, são chamadas de resíduos. O OLS é, portanto, uma forma de estimação de um valor que não existe a partir da relação entre um existente

[1] Há uma polêmica sobre o autor original do OLS. É aceito que Carl Gauss desenvolveu os fundamentos do método de mínimos quadrados no ano de 1795, porém, só publicou essas conclusões em 1809. Antes disso, em 1805, outro matemático, Adrien-Marie Legendre fez uma publicação com os mesmos princípios do método. De qualquer maneira, a literatura reconhece em Gauss a responsabilidade pela proposta do OLS.

e outro teórico, considerando as menores distâncias entre eles, ou seja, os menores resíduos. Este estimador minimiza a soma dos quadrados dos resíduos para maximizar o ajuste do modelo às observações.

No caso da ciência política, poderíamos ilustrar o uso dos princípios do OLS ao considerar que os fenômenos estudados são tão complexos e tão caóticos que o volume de ruídos, se considerarmos as manifestações tais como elas são, seria tão grande que inviabilizaria qualquer análise compreensível. O custo da simplificação feita pelos mínimos quadrados é que, para reduzir os ruídos, perdemos o contraste dos limites entre os casos, ou seja, os componentes dos fenômenos políticos são tratados como se fossem mais homogêneos do que eles realmente são. Para facilitar a compreensão desse efeito, é possível ilustrá-lo com a utilização de filtros de imagem gaussianos no tratamento de fotografias, pois o princípio é o mesmo. A seguir são apresentadas duas cópias de uma mesma imagem. A da esquerda é a original, com ruídos, e a da direita é a que recebeu o filtro gaussiano.

Figura 7.1. Retirada de ruídos de imagem com uso de filtro gaussiano

Fonte: Domínio público, disponível em https://www.mathworks.com/matlabcentral/fileexchange/46866-watermark-dct?ue

Em toda imagem fotográfica os componentes de frequências são proporcionais aos tons de cinza e à distância entre um tom e outro. Uma região homogênea de uma imagem é aquela em que aparecem poucas variações de cinza em largas distâncias.

Já as variações abruptas de cinza geram limites (bordas) de frequências altas, onde se encontram os ruídos. No fundo, a remoção do ruído pelo filtro gaussiano nada mais é que o "aparamento" de bordas agudas em áreas de mudança abrupta de tons de cinza. O que o filtro de Gauss fez na imagem da direita foi manter todos os tons de cinza de baixa frequência e cortar os de alta frequência, os ruídos. Com isso, tem-se a impressão que a imagem ficou mais suave, quando na verdade o que ele fez foi suavizar as zonas de alta frequência nos limites entre tons de cinza.

O mesmo princípio é aplicado aos testes de regressão. Cálculos matemáticos suavizam as diferenças entre as unidades de análise reais, deixando-os mais próximos ou parecidos. Com isso é possível usar valores existentes para propor valores preditos, que não existem na realidade. É importante não esquecer que ao usar o OLS, estamos suavizando as bordas e deixando a imagem real livre de ruídos, porém borrada (*blurring*), o que não se deseja nem no tratamento de fotografias, nem nas análises de regressões de fenômenos políticos. Para evitar análises "borradas", o pesquisador precisa conhecer as limitações da técnica e respeitar os pressupostos das regressões pelo método OLS.

7.1 Começando pelo início: regressão linear simples

A partir do momento em que se estabelece uma correlação entre duas variáveis é possível imaginar ser possível predizer, pelo menos em parte, o comportamento de uma sobre a outra. Ou seja, partindo de valores reais, estabelecer valores teóricos possíveis em função da relação já identificada entre as variáveis – sempre tendo em conta os erros. Regressão é um teste estatístico que existe para fazer predições de valores de uma variável a partir de uma ou mais variáveis (Triola, 1999). Quando se usa apenas uma variável como preditora ou explicativa no teste, ela é chamada de **regressão simples**. Quando há mais de uma preditora, o teste é **regressão múltipla**. Por exemplo, se existir uma correlação entre notas obtidas em um concurso vestibular e desempenho acadêmico na faculdade, é possível usar os resultados do vestibular para prever o desempenho dos alunos na faculdade. Nesse sentido, quanto maior a correlação entre as duas variáveis, mais precisa será a predição, até que para $r^2 = 1$ não haja erro. Por

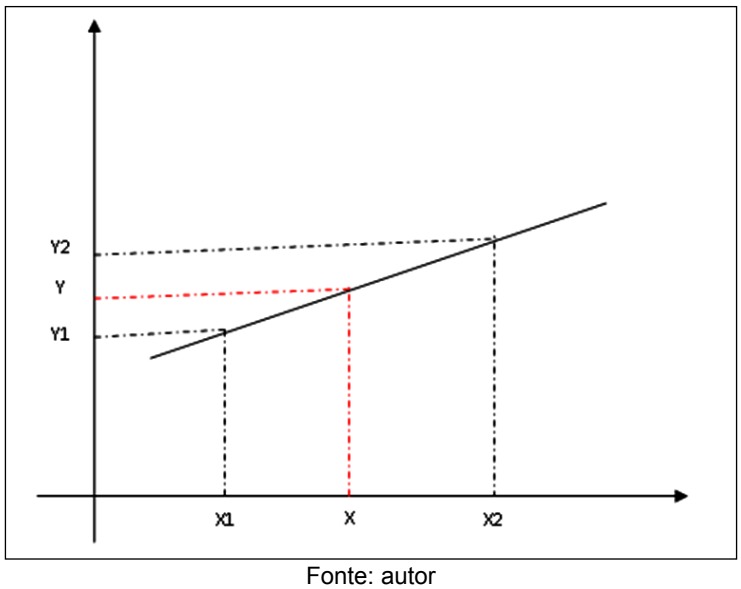
outro lado, quando não é constatada correlação entre as variáveis, não faz sentido tentar encontrar predições entre elas. Predição consiste em estimar valores de uma variável dependente a partir de uma ou mais independentes. Uma regressão linear permite testar quanto duas variáveis estão linearmente relacionadas e a calcular a força dessa relação.

Antes de discutirmos as regressões propriamente ditas, vamos demonstrar de onde vem a lógica por trás de uma equação de regressão pelo método dos mínimos quadrados (OLS) – aquele que considera os vizinhos para "suavizar" as diferenças e prever com o menor erro possível. Trata-se da equação da reta a partir de coordenadas pré-estabelecidas. O gráfico 7.1 mostra a relação entre três pares de valores (y1 e x1; y2 e x2; y_i e x_i) para duas variáveis (x e y). Quando falta informação para uma das variáveis em um dos casos – no gráfico 7.1 perceba as linhas tracejadas em vermelho (y_i e x_i) – parte-se do princípio de que é possível traçar uma reta entre os pontos conectando-os para preencher o valor faltante, que pode ser no eixo x ou y. A fórmula para o cálculo é chamada de equação da reta, como segue abaixo:

$$Fórmula\ da\ Equação\ da\ Reta = \frac{x_i - x1}{x2 - x1} = \frac{y_i - y1}{y2 - y1}$$

Onde, x1 e x2 são os valores conhecidos do eixo horizontal; y1 e y2 são os valores conhecidos do eixo vertical; e x_i e y_i são os valores a serem encontrados.

Gráfico 7.1. Método da reta para identificação de valores



Fonte: autor

Vejamos, por exemplo, a aplicação da equação da reta para o cálculo entre duas escalas de frequência. Frequência 1, na qual o limite é de 90 e frequência 2, na qual o limite sobe para 100. Temos os valores das frequências 1 e 2 para dois alunos (A e B). No entanto, só temos o valor da frequência 1 para o aluno C, como demonstrado no quadro 7.1 a seguir.

Quadro 7.1. Exemplo de distribuição de valores com valor faltante

Aluno	Freq1.(x)	Freq2.(y)
A	60	60
B	90	100
C	80	**?**

Então, temos como pontos das coordenadas, os seguintes:

(60, 60) (90, 100)

Falta o valor de y_i para o Aluno C. Substituindo na fórmula indicada acima, temos que:

$$\frac{x-60}{90-60} = \frac{y-60}{100-60}$$

$$\frac{x-60}{30} = \frac{y-60}{40}$$

$$40.(x-60) = 30.(y-60)$$

$$40x - 2.400 = 30y - 1.800$$

$$40x - 30y - 2.400 + 1.800 = 0$$

$$40x - 30y - 600 = 0$$

$$\boldsymbol{Resultado = 4x - 3y - 60 = 0}$$

Com a equação acima, podemos substituir o valor de x_i pelo conhecido para o aluno C (no caso, o valor 80) e encontraremos o valor faltante (y_i) para o aluno C, como segue:

$$3y = (4x80) - 60$$

$$3y = 320 - 60$$

$$3y = 280$$

$$y = \frac{280}{3} = \mathbf{93,33}$$

Assim, temos que para o aluno C, que apresentou frequência de 80 em uma escala cujo máximo é 90, teria uma frequência de 93,33 no caso da frequência seguir até 100. Essa equação é usada para resolver problemas de variação linear e serve de base para os cálculos que virão a seguir, fundamentados nos princípios da reta de regressão. Evidente que nos casos da ciência política, os valores nunca serão absolutos, pois existe um número grande de variáveis interferindo na relação entre duas características. Assim, é importante lembrar que em todos os cálculos de regressão deve-se considerar o fator de Erro, que é o percentual da variável que não pode ser explicado pela relação entre as variáveis utilizadas na equação. Agora que já conhecemos o princípio da redução das distâncias pelos mínimos quadrados e sabemos que é possível encontrar um valor faltante usando a equação da reta, podemos passar aos fatores que integram a equação da reta da regressão linear.

7.2 Fórmula da Regressão Linear

Uma regressão nada mais é do que a aproximação de uma linha reta que passa por uma nuvem de pontos em um diagrama de dispersão na trajetória que permite a menor distância possível de todos os pontos. Neste caso temos um estimador que é o Melhor Estimador Linear Não enviesado, cuja sigla em inglês é BLUE.

O objetivo dessa reta é sintetizar e representar a nuvem de pontos, podendo ser usada como preditora de valores de uma variável em função da outra (Gujarat, 2006). Para se calcular a reta da regressão linear, são necessários apenas cinco componentes: a média de X, a média de Y, o Desvio Padrão de X, o Desvio Padrão de Y e o coeficiente de correlação (r). Com eles, é possível montar os fatores que compõem

a fórmula da regressão de uma variável Y sobre uma variável X, como descrita abaixo (invertendo dois fatores é possível regredir a variável X sobre a Y):

$$\bar{x} = \alpha + \beta . y + \varepsilon$$

ou

$$\bar{y} = \alpha + \beta . x + \varepsilon$$

Onde:

α(coeficiente linear): é o *intercept*o ou a ordenada de origem e indica a interseção da linha das ordenadas, ou seja, a que altura o eixo Y é *intercept*ado pela reta de regressão. Em outras palavras, o *intercept*o representa o valor de Y quando X é zero. Trata-se de uma constante a ser adicionada para que a média das predições seja igual à média dos valores obtidos. Uma das fórmulas mais simples para obter o valor de α é a que dispensa o uso das medidas de variação. É a que segue:

$$\alpha = \mu Y - (\beta \; x \; \mu X)$$

β* (coeficiente angular ou de regressão)**: o ângulo (*slope*) representa o número de unidades modificadas em Y para cada unidade de mudança de X. Essas mudanças são medidas pelo ângulo da linha que oferece a melhor estimativa linear para Y de X. Em uma regressão múltipla, existem vários coeficientes ***β, um para cada variável independente, que são chamados de coeficientes angulares parciais. O ***β*** expressa a declividade da reta de regressão, sendo calculado a partir da seguinte fórmula:

$$\beta = \frac{\sum X.Y - (n \; x \; (\mu X \; x \; \mu Y))}{\sum X^2 - (n \; x \; \mu X^2)}$$

Outra forma de encontrar o ***β*** é a partir do cálculo que usa o coeficiente de correlação e os desvios padronizados das variáveis. Nesse caso, a fórmula seria:

$$\beta = r \; x \; (\frac{SDy}{SDx})$$

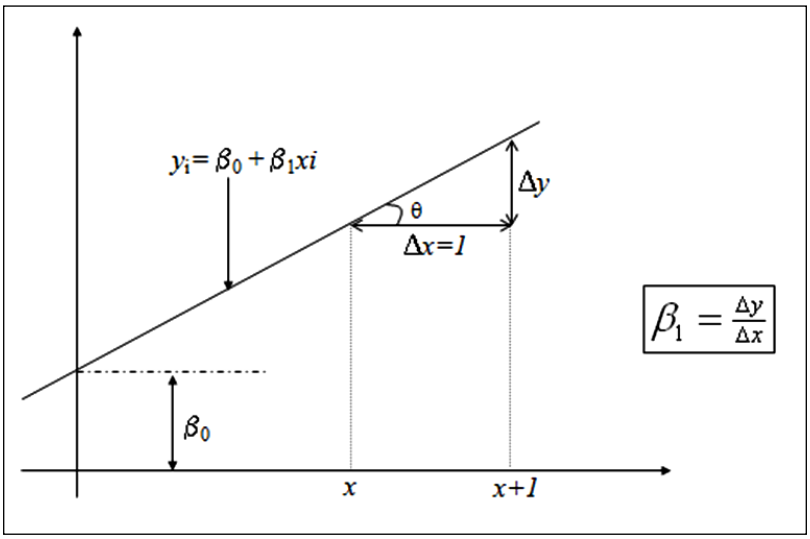
ε (fator de erro): representa o termo de erro aleatório, presente em todas as equações preditivas. O erro aleatório também é chamado de resíduo e pode ser calculado por:

$$\boldsymbol{\varepsilon} = (y - \bar{y})$$

Onde $\bar{y}$ é o valor predito pela reta de regressão e y é o valor observado. Portanto, a diferença entre o predito e o observado é o resíduo ou o fator de erro.

No gráfico 7.2 abaixo, o Beta é dividido em duas partes. O primeiro, representado por β_0 (*intercepto*), que também é denominado de Alfa, ocorre quando a região experimental inclui X = 0. Portanto, β_0 é o valor da média da distribuição de Y em *X* = 0, e ela não tem significado como termo isolado dentro do modelo. O outro Beta é o β_1 (inclinação), que expressa a *taxa de mudança* em *Y*, ou seja, é a mudança em *Y* quando ocorre a mudança de uma unidade em X. Esse é o valor mais importante do modelo, pois é a partir dele que será feito o cálculo de predição. Além disso, ele indica a mudança na média da distribuição de probabilidade de *Y* por unidade de acréscimo em *X*. Portanto, o que nos interessa aqui é o Beta angular.

Gráfico 7.2 – Representação de uma reta de regressão



O resultado de uma equação de regressão gera como principal informação o Beta (B), que é a determinação da explicação sobre a variação da dependente. Ou seja, quanto da mudança na variável independente gera de alteração na variável dependen-

te. Esse coeficiente é acompanhado do grau de significância, sendo considerado estatisticamente significativo o resultado que demonstrar significância abaixo de 0,050. A estatística t, contida nos resultados da equação, também é um indicador de validade da relação. Quanto maior for essa estatística, maior será a determinação de uma variável sobre outra. Quando a regressão é múltipla (existe mais de uma variável independente em relação a uma dependente) torna-se útil o valor de Beta padronizado. Isso porque nem todas as variáveis independentes estarão na mesma unidade. O resultado de Beta padronizado transforma as unidades de cada variável em unidades de desvio-padrão, tornando possível a comparação entre elas para a identificação de qual apresenta maior importância na explicação do fenômeno analisado na variável dependente.

Além das estatísticas descritivas, que tratam dos resultados obtidos na relação entre as variáveis testadas, uma regressão permite produzir estatísticas inferenciais, ou seja, indicadores que permitem fazer predições futuras a partir das relações descritas das variáveis. Os principais resultados preditivos em um modelo de regressão são *t-value* e *p-valor*. Se o objetivo do pesquisador é produzir modelos mais adequados para predição, ele precisa prestar atenção nesses indicadores. Os modelos são mais adequados quando incluem mais variáveis preditoras ou quando incluem variáveis preditoras de maior qualidade. Em uma regressão linear simples, quando se acrescenta uma segunda variável explicativa, a tendência é que haja um aumento na capacidade preditiva do modelo.

Um dos coeficientes produzidos em um modelo de regressão é o coeficiente de correlação múltipla (r), que é a correlação entre valores observados e valores preditos. É possível identificar a intensidade do coeficiente de correlação múltipla gerando-se um gráfico de dispersão entre os valores observados de X e os valores preditos de Y. A forma de distribuição dos valores preditos nesse gráfico de dispersão é um indicador da qualidade preditiva do modelo proposto. Como todo teste estatístico, a regressão linear possui alguns pressupostos que não podem ser quebrados, caso contrário a análise dos resultados fica prejudicada. Os pressupostos são:

- Distribuição Normal da variável dependente (Y), que pode ser verificada através do gráfico QQ e histograma da variável Y (ver capítulo 7);

- Todas as variáveis explicativas devem estar na medida intervalar, de razão ou dicotômica, e a variável dependente deve ser obrigatoriamente contínua, intervalar ou de razão;

- Existência de relação linear entre os casos da variável dependente (Y) e os casos da variável independente (X), que pode ser verificada em um gráfico de distribuição (*scatterplot*);

- Homoscedasticidade, que é a distribuição/variância dos erros de forma homogênea/constante ao longo de toda a reta de regressão;

- Confiabilidade (*reliability*) das variações de X sobre Y;

- A especificação do modelo deve seguir os seguintes passos: a) todos os preditores relevantes devem ser incluídos na análise; e b) os preditores irrelevantes devem ser retirados da análise;

- Além de se distribuírem igualmente entre os conjuntos de variáveis independentes, não deve existir autocorrelação entre os termos de erro das variáveis independentes. Também não deve existir correlação entre os erros e as variáveis independentes;

- No caso de regressão múltipla, não deve existir multicolinearidade entre as variáveis independentes, que é quando se dá uma perfeita combinação linear entre as variáveis independentes;

- Amostra de casos randômica e representativa (Bunchaft & Kellner, 1999).

Para verificar se os pressupostos da relação entre as duas variáveis estão mantidos faz-se uma análise de resíduos $(y - \bar{y})$ com um gráfico de dispersão entre a variável explicativa X no eixo de x e os resíduos no eixo de y. Nesse gráfico, a distribuição dos casos deve ser em forma de nuvem de ponto, indicando ausência de relação entre os resíduos e a variável preditora. Caso esse gráfico de dispersão apresente uma relação evidente entre as variáveis, trata-se de uma quebra de pressupostos.

Aplicaremos os cálculos para encontrar os principais coeficientes de regressão linear para duas variáveis de um banco de dados sobre o desempenho dos partidos nas eleições para deputado federal em 2014. A unidade é o partido político, com N = 32. Usaremos as variáveis “total de votos de legenda para deputado federal” (votleg) e “número de doações de pessoas físicas” (doapfis) para as campanhas dos partidos. O objetivo é realizar os cálculos dos coeficientes de uma regressão linear. Porém, eles podem testar a hipótese de que quanto maior o número de doadores para um partido, maior tende a ser o número de votos de legenda desse partido, já que se trata de dois indicadores de dimensões diferentes (dinheiro e voto) que conectam representantes

e representados. A variável explicativa (X) será número de doadores (aqui é usado o número de operações de doação e não os valores delas, pois sabemos que por falta de um limite real havia grandes distorções entre os partidos quanto aos valores recebidos em doações declaradas). A variável dependente (Y) é o total de votos de legenda obtidos pelo partido. Assim, a hipótese é que quanto mais doadores um partido tiver, maior tenderá a ser o seu número de votos de legenda. Se o modelo apresentar bom ajuste, com baixos erros e alta capacidade preditiva, ele pode ser usado para estimar quanto um partido deve ter de votos de legenda a partir do número de doações que ele recebe durante a campanha. A tabela 7.1 abaixo apresenta as informações das variáveis originais e as novas variáveis necessárias para os cálculos dos coeficientes.

Tabela 7.1 – Doadores de campanha e votos de legenda para deputado federal em 2014

ID	PARTIDO	DOAPFIS (X)	VOTLEG (Y)		X*Y	X^2
1	DEM	618	499.776		308.861.568	381.924
2	PC do B	2.162	237.214		512.856.668	4.674.244
3	PCB	77	64.351		4.955.027	5.929
4	PCO	2	9.117		18.234	4
5	PDT	1.716	953.524		1.636.247.184	2.944.656
6	PEN	515	113.975		58.697.125	265.225
7	PHS	613	121.148		74.263.724	375.769
8	PMDB	4.178	1.609.274		6.723.546.772	17.455.684
9	PMN	314	95.856		30.098.784	98.596
10	PP	1.170	821.703		961.392.510	1.368.900
11	PPL	205	78.116		16.013.780	42.025
12	PPS	1.380	226.246		312.219.480	1.904.400
13	PR	2.033	522.098		1.061.425.234	4.133.089
14	PRB	2.347	383.298		899.600.406	5.508.409
15	PROS	803	210.234		168.817.902	644.809
16	PRP	596	197.128		117.488.288	355.216
17	PRTB	371	70.141		26.022.311	137.641
18	PSB	4.211	1.446.945		6.093.085.395	17.732.521
19	PSC	1.171	274.766		321.750.986	1.371.241
20	PSD	2.401	685.674		1.646.303.274	5.764.801
21	PSDB	3.363	3.748.674		12.606.790.662	11.309.769
22	PSDC	495	56.858		28.144.710	245.025
23	PSL	351	116.647		40.943.097	123.201
24	PSOL	1.877	542.133		1.017.583.641	3.523.129
25	PSTU	208	85.679		17.821.232	43.264
26	PT	12.799	3.848.601		49.258.244.199	163.814.401
27	PT do B	427	74.665		31.881.955	182.329
28	PTB	1.593	478.467		762.197.931	2.537.649
29	PTC	328	75.475		24.755.800	107.584
30	PTN	210	128.065		26.893.650	44.100
31	PV	1.498	465.145		696.787.210	2.244.004
32	SD	2.393	189.765		454.107.645	5.726.449
MÉDIA		1.638	575.971	SOMA	85.939.816.384	255.065.987

Fonte: autor a partir de dados do TSE

O primeiro passo dos cálculos indicados por (Bunchaft & Kellner, 1999) é obter o valor de β como segue:

$$\beta = \frac{\sum X.Y - (n\ x\ (\mu X\ x\ \mu Y))}{\sum X^2 - (n\ x\ \mu X^2)} = \frac{85.939.816.384 - (32\ x\ (1.638{,}28\ x\ 575.971{,}18)}{255.065.987 - (32\ x\ 1.638{,}28^2)}$$
$$= \frac{85.939.816.384 - 30.195.266.072{,}65}{255.065.987 - 85.886.763{,}46} = \frac{55.744.550.311{,}35}{169.179.223{,}54}$$
$$= \mathbf{329{,}49}$$

O segundo passo é calcular o valor de α para a regressão, conforme a fórmula que segue:

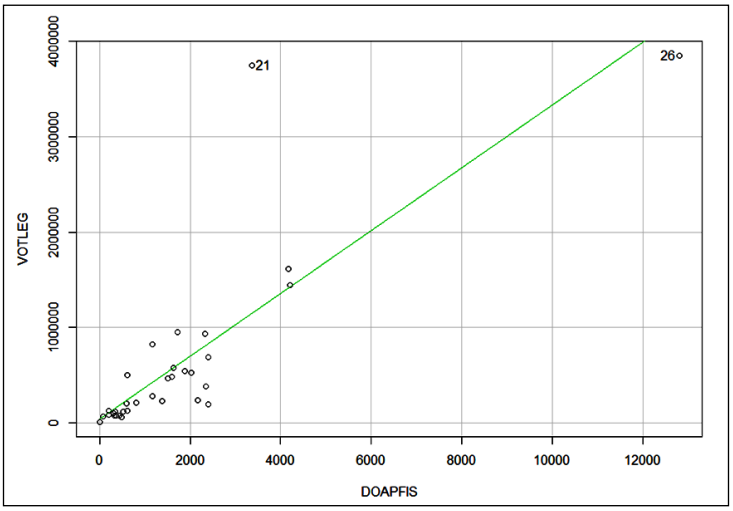
$$\alpha = \mu Y - (\beta\ x\ \mu X) = 575.971{,}18 - (329{,}49\ x\ 1.638{,}28) = 575.971{,}18 - 539.796{,}87$$
$$= \mathbf{36.174{,}31}$$

Agora, com todos os termos da equação conhecidos, já podemos calcular os valores preditos de Y a partir de valores existentes de X. Digamos que se encontrar o valor predito de votos de legenda para um partido que obteve 1.600 doações de pessoas físicas, basta substituir os valores na formula da reta:

$$\bar{y} = \alpha + \beta\ .x + \varepsilon = 36.174{,}31 + (329{,}49\ x\ 1.600) + \varepsilon = 563.358{,}31 + \varepsilon$$

Assim, sabemos que o valor predito de votos de legenda para um partido que obteve 1.600 doações de pessoas físicas é de 563,3 mil mais o erro. Os dados originais do exemplo e as representações da reta de regressão estão demonstrados no gráfico de dispersão a seguir. Para melhor visualizar a relação entre as duas variáveis usadas no exemplo aqui, o gráfico 7.3 indica a distribuição dos valores das duas variáveis.

Gráfico 7.3. Reta de regressão para doações de pessoas físicas e votos de legenda



Fonte: autor a partir de TSE.

No *RCommander* é possível solicitar a marcação de casos extremos. Os dois pontos marcados no gráfico 7.3 indicam os partidos que ficaram mais fora do padrão dos demais. O caso número 21 é o PSDB, que teve muitas doações de pessoas físicas, porém um número muito maior de votos de legenda que os demais partidos. Já o caso 26 é o PT, que teve o maior número de operações de doações de pessoas físicas e o maior número de votos de legenda para deputado federal em 2014, ainda que tenha ficado abaixo da reta de regressão. Para rodar o teste de regressão linear no *RCommander* o caminho é "Estatísticas/Ajuste de Modelo/Regressão Linear". Os resultados da regressão entre as duas variáveis apresentadas acima seguem no quadro:

```
Linha de comando: Rcmdr>  RegModel.1 <- lm(VOTLEG~DOAPFIS,
data=legenda)

Resultados: Rcmdr>  summary(RegModel.1)
Call: lm(formula = VOTLEG ~ DOAPFIS, data = legenda)

Residuals:
    Min      1Q  Median      3Q     Max
-638886 -139120  -69944    -928 2599925

Coefficients:
             Estimate Std. Error t value      Pr(>|t|)
(Intercept)  38966.33  109757.01   0.355         0.725
DOAPFIS        330.00      39.45   8.365 0.00000000147 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 513800 on 32 degrees of freedom
Multiple R-squared:  0.6862,  Adjusted R-squared:  0.6764
F-statistic: 69.98 on 1 and 32 DF,  p-value: 0.000000001473
```

A primeira informação que o *RCommander* fornece é da distribuição dos resíduos, com valores mínimo, máximo e quartis de resíduos (Freijo, 2013). Como veremos a seguir, os resíduos são importante capítulo da análise de regressão. Em seguida, aparecem os coeficientes individuais, com o (*intercept*) sendo o α (38.966,33) e a estimativa para a variável explicativa "doapesf" sendo o *β* (330,00). Perceba que os valores são um pouco diferentes dos calculados anteriormente, o que é consequência dos arredondamentos realizados na calculadora. A interpretação desses coeficientes é que em um modelo teórico de regressão um partido sem nenhuma doação de pessoa física teria 38.966,33 votos de legenda, pois esse é o ponto em que a reta cruza o eixo Y quando o valor de X é zero. E, para cada doação de pessoa física que o partido recebe, ele passa a ter 330 votos de legenda a mais. Este valor equivale ao ângulo, ou coeficiente angular da reta.

Em seguida, a saída de resultados indica o nível de significância das duas estatísticas. No caso o α não é significativo (*p-valor* = 0,725), mas o coeficiente da variável "doapesf" é altamente significativo, ficando com significância acima de 99% (0,001). Esses indicadores são mais importantes para modelos de regressão múltipla do que de regressão simples. Por fim, aparecem as estatísticas do modelo como um todo. O erro padrão residual (513.800), os graus de liberdade (32), o coeficiente de determinação r^2= 0,686 e o r^2 ajustado= 0,676, além das estatísticas do teste de diferença de médias F=69,98 (*p-valor* = 0,000). O r^2 indica um ajustamento de modelo de 68,6%, ou seja, o crescimento no número de operações de doações de pessoas físicas explica mais de dois terços do crescimento dos votos de legenda.

Um capítulo à parte nos cálculos de regressão é a análise dos erros ou resíduos, devido a importância que essa informação tem para a compreensão dos modelos. Este é o assunto do próximo tópico.

7.3 Erro da reta de regressão (RMS) – análise de resíduos

Este é um bom momento para recordarmos a imagem 7.1 do início do capítulo. Lembre-se que na associação que fiz sobre o uso do método OLS para suavizar ruídos

em imagens, o papel da equação era cortar os picos de contraste de cinza nos limites entre um tom e outro. Com isso o ruído diminuiria, assim como os limites entre diferentes tons. Naquele momento, chamei atenção para o problema de a imagem ficar "borrada" quando se corta em excesso os limites entre espaços com tons distintos. Outro problema é quando temos um tom muito diferente do outro. O pico de tom de cinza é tão grande que a fórmula não consegue cortá-lo o suficiente e o que acontece é que esse tom, que pode ser mais escuro, acaba interferindo em todos os espaços ao seu redor, passando a impressão – errônea – de que naquele espaço os tons de cinza são mais escuros do que o são na verdade. Aplicando à análise política, os tons de cinza discrepantes são os casos muito distintos que podem existir em um banco de dados. Se os valores de alguns casos estão muito fora da norma dos demais (os chamados *outliers*) existirá uma chance real de eles interferirem artificialmente nos resultados dos modelos, gerando efeitos que vão encobrir as verdadeiras relações entre as variáveis. Em outras palavras, se uma variável possui casos discrepantes e em uma regressão eles não são normalizados, esses *outliers* acabam impactando nos coeficientes. Nós conseguimos identificar se existem ou não casos discrepantes a ponto de interferir nos resultados de uma regressão a partir da análise dos resíduos.

Começando pelo exemplo acima, a análise de resíduos nos permite verificar como estão distribuídas as diferenças entre o valor real e o valor estimado pela reta de regressão. Para o modelo estar bem ajustado, principalmente quando o objetivo é fazer inferências, é preciso que os resíduos estejam distribuídos aleatoriamente no modelo (Gujarat, 2006). Para tanto, o ideal é que a mediana dos resíduos esteja muito próxima a zero, o que indicaria metade de resíduos positivos e outra metade negativos. Além disso, os valores máximo e mínimo indicando distâncias equivalentes de zero, apenas com sinais trocados, é outro indicador de que os resíduos estão bem distribuídos. Analisando a saída de resultados acima, percebemos que os resíduos do modelo não estão bem ajustados. A mediana está muito distante de zero (-69.944) e o sinal negativo indica que mais da metade dos resíduos fica abaixo da linha de regressão. Além disso, o valor máximo +2.599.925 fica bem mais distante da mediana que o valor mínimo, que é de -638.886. Ou seja, temos resíduos concentrados entre o valor mínimo e a mediana e dispersos entre a mediana e o valor máximo.

Como estamos estimando o valor de uma variável a partir da reta de regressão, e como quase nunca essa reta equivale a r = ±1, é preciso identificar o erro dessa estimativa. Para isso existe o "rms", que é uma espécie de medida de dispersão para a relação entre duas variáveis. Quando estamos analisando o comportamento de uma única variável a partir de uma curva normal do histograma, a média é usada como medida de tendência central e o Desvio Padrão como medida de variabilidade. Mas, quando analisamos os efeitos conjuntos de duas variáveis a partir de um gráfico de dispersão, a reta da regressão equivale à medida de tendência central e o erro da reta de regressão (rms) equivale à medida de dispersão. A fórmula para calcular o intervalo do "rms" de duas variáveis em uma regressão linear é a seguinte:

$$rms = \sqrt{1 - r^2}\ x\ DPy$$

Para obter o desvio padrão de votos de legenda (DPy) no *RCommander* o caminho é "Resumos/Resumos numéricos" e seleciona Desvio Padrão. No caso do exemplo acima, o "rms" seria:

$$rms = \sqrt{1 - (0{,}686)^2}\ x\ 903240{,}6 = 0{,}727\ x\ 903240{,}6 = 656.655{,}91$$

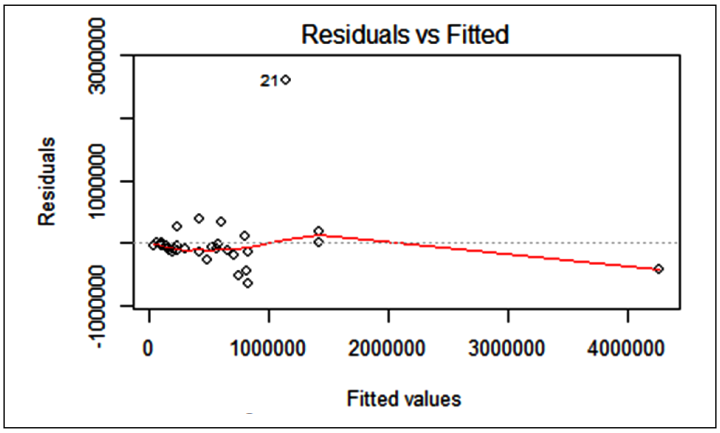
Isso significa que uma unidade rms acima da reta de regressão representará o ponto na reta + 656.655,91 pontos, enquanto uma unidade abaixo da reta de regressão representará o ponto da reta – 656.655,91 pontos. Em uma regressão onde os erros são distribuídos aleatoriamente, ± 1 rms da reta de regressão incluem 68% dos casos e ± 2 rms da reta de regressão incluem 95% dos casos. Uma regressão linear entre duas variáveis deve apresentar as seguintes características:

- a média dos resíduos deve ser igual a zero;
- Não deve existir associação linear entre os resíduos de Y e a variável X;
- Em um gráfico de resíduos não deve existir nenhuma tendência linear.

Se não houver correlação linear (r = 0), o rms será igual ao desvio padrão de Y. Se os erros forem aleatórios, os resíduos do rms não poderão formar tendência. Em um gráfico de distribuição, será formada uma nuvem de pontos. Os resultados de um teste de regressão só podem ser levados em consideração se os resíduos forem distribuídos aleatoriamente, ou seja, apresentarem independência entre si. É sempre recomendável

analisar o gráfico de resíduos antes de rodar a regressão linear.. No *RCommander* o gráfico de resíduos de uma regressão é obtido em "Modelos/Gráficos/Diagnósticos Gráficos Básicos". Serão gerados quatro gráficos de diagnósticos, porém, por agora só nos interessa o primeiro, que está reproduzido a seguir.

Gráfico 7.4. Distribuição dos resíduos da regressão



Analisando a imagem é possível perceber que a maior parte dos casos tendeu a se distribuir aleatoriamente em torno de zero, porém existem dois resíduos que ficam muito fora da distribuição geral. Já vimos no gráfico anterior (gráfico 7.2) quais são esses casos: PSDB e PT. São eles que estão criando uma tendência na distribuição dos resíduos. Os resíduos oferecem informações sobre o comportamento geral do modelo, assim como sobre os casos individuais que estão gerando influências e possivelmente distorcendo os resultados.

7.4 A ESTATÍSTICA "T" E OS TESTES COMPLEMENTARES DE AJUSTAMENTO DO MODELO

A estatística t é um coeficiente de teste de diferença de médias, portanto, no teste de regressão, ela tem a função de um teste de hipóteses. Aqui, a hipótese a ser testada é se a variável X tem interferência sobre as variações de Y. Então, teríamos as hipóteses:

H_0 = Não há relação entre X e Y (que equivale a $\beta 1 = 0$).

H_1 = Há relação entre X e Y (que equivale a $\beta 1 \neq 0$).

Se β_1 for igual a zero, então a equação equivale ao *intercept*o mais o erro, logo, não há interferência das variações de X sobre Y. A reta não apresentaria ângulo nesse caso e teríamos uma dispersão em forma de nuvem de ponto, sem direção definida. Ou seja, X não explicaria a variação de Y. Mas, se β_1 for diferente de zero a hipótese nula é rejeitada e o coeficiente indica que as variações de X têm alguma influência sobre as variações de Y. Nesse caso, rejeitamos a hipótese nula. O Coeficiente t serve para nos indicar quão longe de zero está β_1, pois ele pode estar próximo o suficiente de zero para não ser considerado estatisticamente significativo. Nesse caso, não se pode rejeitar a hipótese nula. Em resumo, o coeficiente "t" serve para nos indicar se o β_1 está longe o suficiente de zero para aceitarmos alguma dependência nas variações de X e Y. Ele serve para indicar a precisão do modelo. Quanto mais longe de zero, mais precisa será a predição do modelo quanto à associação entre X e Y (Triola, 1999). Na prática, o coeficiente t é calculado a partir da seguinte fórmula:

$$t = \frac{\beta 1}{SE(\beta 1)}$$

Onde,

β_1 = coeficiente angular;

SE = erro padrão médio;

O resultado será a distância de β_1 em relação a zero em número de desvio padrão. Aplicando para os resultados do exemplo utilizado até aqui, a saída de resultados da regressão indica como β1 da variável "doapfis" 330,00 e o erro padrão médio dessa variável é 39,45. Assim, temos que:

$$t = \frac{\beta 1 - 0}{SE(\beta 1)} = \frac{330}{39{,}45} = 8{,}36$$

Para até N = 30, os coeficientes t > 2 indicam distância suficientemente grande em relação a zero e, portanto, diferenças estatisticamente significativas que permitem rejeitar a hipótese nula. No nosso exemplo, o N = 32 e o t = 8,36 pode ser considerado estatisticamente significativo. Sendo assim, podemos afirmar que há um alto grau de predição da variável independente sobre a dependente. Isso também nos é informado

pelo valor de *p-valor* que consta no quadro de saída da regressão.

Apesar do teste de regressão linear ser muito útil e de fácil interpretação, são necessários alguns cuidados para a sua aplicação. Existem várias, mas as principais restrições à regressão podem ser resumidas em:

- Heteroscedasticidade: Ocorre quando há diferenças nas variâncias dos resíduos em relação à reta. O ideal é que essas diferenças de resíduos sejam constantes (homoscedásticas). Se o modelo inclui apenas uma variável independente é fácil perceber a heteroscedasticidade no gráfico de dispersão, pois as distâncias entre os pontos e a reta tendem a aumentar ou diminuir. Para testar a existência de heteroscedasticidade na regressão, o *RCommander* oferece o teste de *Breusch-Pagan*. Ele é baseado no teste do χ^2 considerando os graus de liberdade da regressão. Se para um intervalo de confiança de 95% o *p-valor* ficar abaixo de 0,050 então a hipótese nula de homoscedasticidade é rejeitada e aceita a existência de distribuições heteroscedásticas dos resíduos.

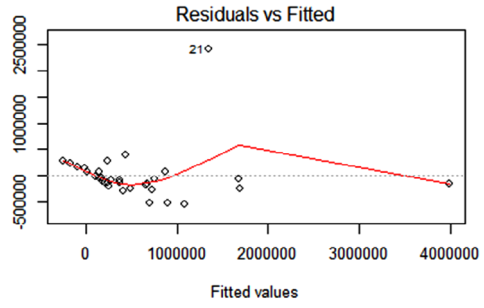
- Multicolinearidade: Quando a regressão conta com mais de uma variável independente e entre elas existem altos coeficientes de correlações cruzadas, induzindo a um viés. O teste para verificar a existência de multicolinearidade é o VIF (*Variance Inflating Factor*) que só deve ser aplicado em regressões múltiplas por motivos óbvios. Se o valor de VIF fica abaixo de 10,0 significa que não existe colinearidade entre as variáveis explicativas do modelo.

- Cointegração: Ocorre quando não há independência entre os dados por eles estarem distribuídos ao longo do tempo, gerando uma tendência temporal, o que rompe o pressuposto da independência entre as observações. Isso se aplica apenas aos casos em que os dados foram tomados em diferentes momentos do tempo. Para identificar a cointegração, existem técnicas específicas de análises de séries temporais, pois é preciso levar em conta a ordem de coleta dos dados. A estatística mais usada para medir cointegração em dados com dependência temporal é a *Durbin-Watson*. As análises de séries temporais serão apresentadas em outro capítulo.

Para testar a heteroscedasticidade e multicolinearidade, agregaremos outra variável ao modelo de regressão usado como exemplo aqui. Vamos usar o número de candidatos a deputado federal por partido para tentar explicar o número de votos de legenda. Assim, o modelo que antes era simples, agora passa a ser múltiplo, com “número

de doadores pessoa física" e "número de candidatos" usados para explicar as variações no número de votos de legenda. O quadro a seguir mostra a saída da regressão com todos os coeficientes já discutidos anteriormente e os testes de heteroscedasticidade e de colinearidade. O teste de heteroscedasticidade é obtido no *RCommander* seguindo o caminho: "Modelos/Diagnóstico Numérico/Teste *Breusch-Pagan* para heteroscedasticidade". Para o teste de multicolinearidade o caminho é: "Modelos/Diagnóstico Numérico/ Fatores de Inflação de Variância".

```
Linha de Comando da regressão múltipla
Rcmdr>  summary(RegModel.2)

Call:lm(formula = VOTLEG ~ DOAPFIS + NUMCAND, data = legenda)

Resultados:
Residuals:
    Min      1Q  Median      3Q     Max
-536198 -184868  -81185   82254 2406319

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -285052.25  201150.90  -1.417    0.1671
DOAPFIS         266.83      51.03   5.229 0.0000134 ***
NUMCAND        2818.78    1476.11   1.910    0.0661 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 508200 on 29 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7205,  Adjusted R-squared:  0.7012
F-statistic: 37.38 on 2 and 29 DF,  p-value: 0.000000009391

Linha de commando do teste de heteroscedasticidade
Rcmdr>  bptest(VOTLEG ~ DOAPFIS + NUMCAND, varformula = ~ fitted.
values(RegModel.2), studentize=FALSE, data=legenda)

Resultados:
      Breusch-Pagan test
data:  VOTLEG ~ DOAPFIS + NUMCAND
BP = 10.007, df = 1, p-value = 0.001559
```




```
Linha de commando do teste de multicolinearidade:
Rcmdr>  vif(RegModel.2)

Resultados:
 DOAPFIS  NUMCAND
1.705507 1.705507
```

Sobre as estatísticas do modelo múltiplo, quando comparadas com o exemplo anterior, percebe-se um crescimento no ajuste, com r² passando a 0,720. Em relação às estatísticas individuais, apenas a variável "número de doadores pessoa física" tem efeito estatisticamente significativo sobre a variação de votos de legenda. A variável "número de candidatos" apresenta t = 1,910 (*p-valor* = 0,066), portanto, não deve ser considerada significativa.

Os testes de ajustamento mostram os seguintes resultados: O *Breusch-Pagan* para heteroscedasticidade apresenta coeficiente 10,00 e *p-valor* = 0,001, considerando intervalo de confiança de 95%, o *p-valor* fica abaixo do limite crítico o que nos obriga a rejeitar a hipótese de homoscedasticidade. Os resíduos não se distribuem de maneira aleatória. O gráfico de distribuição dos resíduos adicionado à saída de resultados acima indica que mesmo na regressão múltipla os resíduos de PSDB e PT continuam gerando tendência. Sobre a multicolinearidade, o VIF apresentou valor de 1,705. Como são apenas duas variáveis, o coeficiente será o mesmo. Se houvesse uma terceira, os valores mudariam. O importante aqui é identificar que o VIF ficou muito abaixo de 10,0, portanto, não sendo constatada multicolinearidade.

Quando o pesquisador possui dois modelos distintos e precisa compará-los para saber qual o melhor ajustamento, isso não pode ser feito pelo r² – pois eles terão números de variáveis e dimensões distintas. Para esses casos, existem os testes AIC (*Akaike Information Criteria*) e BIC (*Bayesian Information Criteria*). Eles são complementares, portanto, precisam ser analisados sempre em pares. Quanto menores os valores dos coeficientes de AIC e BIC, mais ajustado estará o modelo em comparação ao outro (Gujarat, 2006). Apenas a título de exemplo, seguem abaixo os valores AIC e BIC para o modelo de regressão linear simples e o modelo múltiplo, com as duas variáveis explicativas. Perceba que para os dois critérios o modelo 2, o da regressão múltipla, é melhor ajustado que o primeiro.

```
Resultados de AIC e BIC para regressão linear simples (Modelo 1)
Rcmdr>  AIC(RegModel.1)
[1] 994.5999
Rcmdr>  BIC(RegModel.1)
[1] 999.179

Resultados de AIC e BIC para regressão linear simples (Modelo 2)
Rcmdr>  AIC(RegModel.2)
[1] 936.5404
Rcmdr>  BIC(RegModel.2)
[1] 942.4033
```

Testes de regressão são úteis quando se pretende fazer predições de valores para relações entre pontos de variáveis a respeito dos quais não se possui informação. Até aqui, estudamos o modelo básico de análise de regressão, que é a linear simples, cujo princípio é que a variável dependente, aquela que será testada, precisa ser obrigatoriamente escalar ou contínua. Mas existem outros modelos nos quais podem ser utilizadas variáveis categóricas dicotômicas ou politômicas. No próximo tópico, discutiremos os princípios da regressão binária, quando a variável dependente assume apenas dois valores, presença ou ausência de determinada característica.

7.5 Regressão Binária Logística

O modelo de regressão linear é aplicado a variáveis dependentes contínuas. Para quando a variável dependente é categórica binária, não é possível usar o mesmo modelo estatístico. Para esses casos, o teste indicado é a regressão logística binária, no qual a variável dependente é dicotômica e as variáveis independentes ou preditoras podem ser categóricas ou contínuas (Menard, 1995). Se a variável dependente for categórica, mas não binária, o teste indicado é o de regressão logística multinomial, que não será tratado aqui. A seguir, apresento apenas as principais características da regressão para casos de variável dependente binária. A fórmula da regressão logística binária em Bunchaft e Kellner (1999) é:

$$\overline{Y} = B_0 + (B_k \, x \, X_k)$$

Onde:

$\overline{Y}$ = valor predito da variável Y (dependente);

B_0 = valor predito em Y quando o valor de X é zero;

B_k = coeficiente de regressão não padronizado;

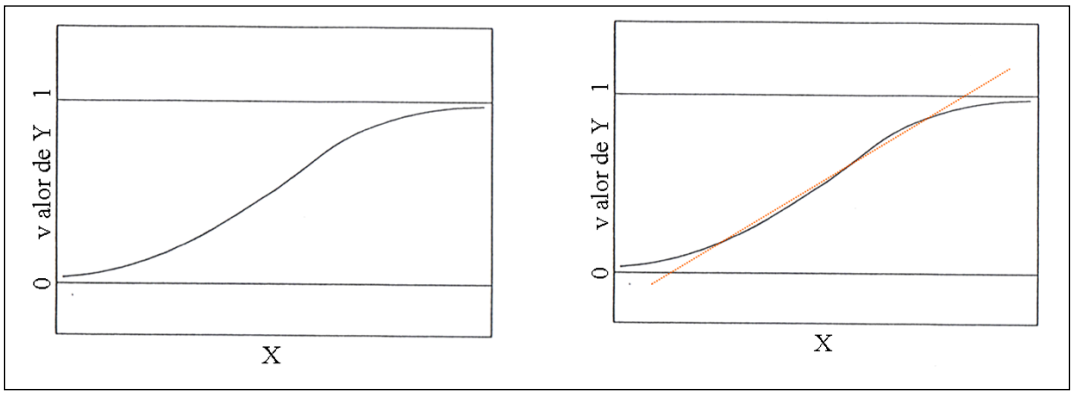
X_k = variáveis preditoras;

K = número de variáveis preditoras;

(Y - $\overline{Y}$) = resíduo ou erro predito.

Na regressão binária, o valor predito pode variar entre zero ou um. Assim, não podemos estabelecer o impacto da variação da variável explicativa sobre a dependente binária, pois não há garantia de combinação linear entre as variações de zero a um, já que o valor estimado de Y não pode ser exatamente igual à média do valor de Y para todos os casos (Menard, 1995), conforme demonstrado no gráfico 7.5 abaixo.

Gráfico 7.5. Representação de valores na regressão binária e linear



A imagem da esquerda do gráfico 7.5 mostra que a relação entre X e Y não é linear, formando uma curva em S, o que viola o pressuposto básico da regressão linear. As principais diferenças são notadas nas extremidades da distribuição. Como se percebe na imagem da direita, ao traçar uma reta de regressão linear nota-se que os valores das extremidades estariam mais distantes da curva em S do que os da parte central. Isso nos obriga a desconsiderar o modelo linear quando a variável dependente for binária e, também, a não considerar mais o efeito em uma unidade de Y em função da mudança de uma unidade de X.

O modelo logístico é um modelo linear para as chances de ocorrência, ou seja, sobre as chances relativas de estar em uma das duas posições possíveis da distribuição. Aqui, o que importa são os *odds* de ocorrência no teste. O logaritmo natural dos *odds* é chamado de *log-odds* ou apenas *logit*. O log diz em quanto aumentam as chances relativas de passar de zero para 1 da variável Y quando se aumenta uma unidade da variável X e o Beta dá o percentual de incremento de valor para a variável dependente. O teste de hipóteses em uma regressão binária logística tem o objetivo de identificar se um preditor individual apresenta efeito significativo ou não na mudança da variável

dependente. Além disso, ele pode indicar se o modelo geral é significativo e, na comparação entre dois modelos, se o modelo A é melhor que o modelo B ou não.

Para testar cada preditor individualmente usam-se os coeficientes de regressão e as razões de odds. Já o teste *Wald* verifica o modelo todo frente a um modelo sem um preditor específico. A questão principal em uma regressão binária logística é como classificar e interpretar os casos corretamente. Entre as estatísticas apresentadas em um modelo de regressão binária logística está o Fator de Inflação da Variância (*VIF*), já discutido acima, que indica quanto a presença de uma variável explicativa no modelo impacta sobre as demais. No caso da regressão binária, a forma mais comum de correção do modelo é pela exclusão de uma das variáveis que apresentam colinearidade. Existem dois testes que indicam variações colineares entre variáveis independentes. Um deles é o teste de Tolerância, cujo resultado precisa estar acima de 0,100 para as variáveis não serem colineares. O outro é o próprio *VIF*, que deve ficar acima de 10 em todas as variáveis para o modelo ser adequado. Quando a variável tem coeficientes que ficam fora dos limites apresentados acima, deve ser retirada do modelo.

Os resultados de uma regressão binária logística sempre expressam a probabilidade de ocorrência de um dos valores preditos de uma variável dicotômica (presença/ausência) e são interpretados em termos de grau de probabilidade de ocorrência. Ou seja, quanto aumenta ou diminui a chance de ocorrência de determinado fato, dado o aumento ou redução de uma unidade na variável independente. Uma estatística importante produzida no modelo de regressão logística binária com mais de uma variável explicativa é a *overall*, que indica qual seria o efeito sobre o modelo, caso uma variável fosse excluída. O *overall* do modelo equivale a um coeficiente χ^2 para a relação entre duas variáveis. Ele indica se os efeitos conjuntos das variáveis explicativas são estatisticamente significativos. Esta estatística é fundamental para os casos em que a regressão pretende fazer predições. Nesses casos, quando o *p-valor* ou nível de significância do *overall* fica acima do limite crítico de 0,050, o modelo não é robusto o suficiente para permitir predições e as análises não devem seguir adiante. Outra estatística usada para medir o impacto individual das variáveis no modelo é o *score ROA*. Trata-se de uma medida de eficiência individual na relação entre a variável explicativa e a dependente. O *ROA* indica o efeito que a ausência da variável geraria caso fosse retirada do modelo.

Não tem limites pré-estabelecidos. Quanto maior o *ROA*, mais efeito individual tem essa variável sobre a dependente.

Em uma regressão binária, os coeficientes de *Cox & Snell* e *Nagelkerke* equivalem ao r² no modelo de regressão linear e indicam quanto das variações totais é explicado pelo modelo. Se o objetivo é buscar por predição de comportamentos, espera-se que o modelo esteja bem ajustado, ou seja, que o coeficiente apresente valor elevado. Trata-se de um indicador para a variância explicada (Menard, 1995). Quando multiplicado por 100, pode ser lido como percentagem.

Feitas as análise do ajustamento do modelo como um todo, podemos passar para as estatísticas individuais das variáveis explicativas, pois na maioria das vezes o que mais interessa em um modelo de regressão é identificar o impacto individual de cada variável independente sobre a dependente. Nesse caso, o primeiro coeficiente individual é o *Wald*, que informa quanto o coeficiente *β* está se distanciando de zero a partir de uma distribuição χ^2. Quanto maior o *Wald*, mais distante de zero estará o *β* e maior a contribuição dessa variável para a mudança de categoria da variável dependente. Em modelos preditivos o nível de significância (*p-valor*) da estatística é importante para demonstrar se os resultados daquela variável independente podem ser usados em predições para toda a população. O quadro 7.2 a seguir apresenta um resumo com descrição das principais estatísticas em um modelo de regressão binária logística.

Quadro 7.2. Principais estatísticas de um modelo de regressão binária logística

Estatísticas do modelo	VIF	Mede o efeito da colinearidade no resultado.
	Tolerância	Identifica colinearidade entre as variáveis.
	Overall (sig.)	Indica os efeitos das variáveis para o modelo.
	ROA	Indica o efeito da ausência da variável para o modelo.
	r² Cox & Snell	Equivale a um coeficiente de determinação do modelo.
	Nagelkerke	Equivale a um coeficiente de determinação do modelo.
Estatísticas individuais	*Wald (β)*	Indica o efeito individual da variável pela distância do β em relação ao zero.
	expβ	Indica o valor esperado de β.
	Odds ratio	A partir de (expβ - 1)*100 identifica o percentual de chance de mudança na variável dependente a partir da mudança de uma unidade na variável independente.

Fonte: autor

Na regressão binária, a estatística individual mais importante é a *ratio odds*, calculada a partir da expectativa de *β* (*expβ*). A *ratio odds* indica qual o impacto individual da variável independente sobre a dependente. Ela é calculada por (Oddsratio=*expβ*-1) e se multiplicarmos o resultado por 100, teremos o resultado que é uma probabilidade em percentual. Tanto aqui quanto no *Wald*, os valores indicam o tamanho do impacto individual e o sinal (positivo ou negativo), a direção do impacto. Um coeficiente com sinal positivo mostra que o impacto é favorável à mudança de categoria da variável dependentes. Já um coeficiente com sinal negativo indica que a relação é desfavorável, ou seja, aquela variável independente contribui negativamente para o efeito esperado.

No *RCommander* a regressão logit é obtida pela produção de modelos lineares generalizados (MLG). O caminho é: "Estatísticas/Ajuste de modelos/Modelo Linear Generalizado". Uma vez na caixa de seleção de variáveis, seleciona-se primeiro a variável dependente com duplo clique sobre ela (irá automaticamente para a caixa antes do sinal til [~]) e depois selecionam-se as variáveis explicativas. Ainda é preciso selecionar o tipo de teste e a função de ligação. No nosso caso, será binominal e logit. No banco de dados usado até aqui, há uma variável binária que separa 1 = partido grande e 0 = partido pequeno. A definição foi dada a partir da mediana do total de votos obtidos para deputado federal (nominais e de legenda) em 2014. Partidos acima da mediana de votos totais são partidos grandes e abaixo, partidos pequenos. Agora, como variáveis explicativas vamos usar: i) o número de doações de pessoas físicas; ii) número total de operações de doações, que inclui todas as formas declaradas na prestação de contas ao TSE; e iii) o total de candidatos a deputado federal apresentados pelo partido. O objetivo do teste é verificar quanto cada uma das variáveis explicativas contribui individualmente para que um partido passe de pequeno para grande no modelo. A saída de resultados está no quadro abaixo:

```
Linha de commando:
Rcmdr>  GLM.14 <- glm(PARTGRD ~ DOAPFIS + NUMCAND + DOATOT,
family=binomial(logit), data=legenda)

Linha de resultados:
Rcmdr>  summary(GLM.3)

Call: glm(formula = PARTGRD ~ DOAPFIS + NUMCAND + DOATOT, family =
binomial(logit), data = legenda)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.29375  -0.47454  -0.05277   0.41915   2.24668
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.296768   2.120846  -2.497   0.0125 *
DOAPFIS     -0.001007   0.002014  -0.500   0.6170
NUMCAND      0.004188   0.010929   0.383   0.7016
DOATOT       0.001943   0.001180   1.646   0.0997 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.361  on 31  degrees of freedom
Residual deviance: 18.729  on 28  degrees of freedom
AIC: 26.729

Number of Fisher Scoring iterations: 7

Rcmdr>  exp(coef(GLM.14))  # Exponentiated coefficients (“odds
ratios”)
(Intercept)     DOAPFIS     NUMCAND      DOATOT
 0.00500775  0.99899342  1.00419684  1.00194439

Rcmdr>  Confint(GLM.3, level=0.95, type=”Wald”)
             Estimate 2.5 %       97.5 %    exp(Estimate)  2.5 %       97.5%
(Intercept) -5.296768 -9.4535504 -1.139986  0.00500        0.00007841  0.3198233
DOAPFIS     -0.001007 -0.0049543  0.002940  0.99899        0.99505787  1.0029445
NUMCAND      0.004188 -0.0172315  0.025607  1.00419        0.98291604  1.0259384
DOATOT       0.001942 -0.0003703  0.004255  1.00194        0.99962967  1.0042645

Rcmdr>  vif(GLM.3)
 DOAPFIS  NUMCAND   DOATOT
5.228265 1.111532 5.318193

Rcmdr>  round(cov2cor(vcov(GLM.14)), 3) # Correlations of parameter
estimates
            (Intercept) DOAPFIS NUMCAND DOATOT
(Intercept)       1.000   0.424  -0.517 -0.604
DOAPFIS           0.424   1.000  -0.015 -0.889
NUMCAND          -0.517  -0.015   1.000 -0.131
DOATOT           -0.604  -0.889  -0.131  1.000
```

Os primeiros resultados são dos desvios de resíduos, que se apresentam com mediana próxima a zero e as distâncias entre máximo e mínimo também próximas de zero, o que indica que os resíduos não devem estar enviesados. Nos coeficientes individuais, o valor da Estimativa equivale ao *Expβ* de cada variável e o seu *p-valor* para cada explicativa. No exemplo, nenhuma variável foi estatisticamente significativa, pois todos os *p-valor*es > 0,050, o que impediria a extrapolação dos resultados de uma amostra para toda população. Em relação aos sinais dos estimadores, a variável número de doações de pessoas físicas apresenta sinal negativo. Isso significa que os partidos com maior número de doações de pessoas físicas tendem a estar na categoria de partidos pequenos. As outras duas variáveis apresentam estimativas positivas, indicando que o aumento nos valores delas contribui para a passagem de partido pequeno para grande.

A questão agora é medir a contribuição comparativa de cada variável explicativa para a dependente. Para isso usamos a fórmula de odds ratio -1 e se o resultado for multiplicado por 100 podemos ler em termos percentuais. A variável "doações de pessoas físicas" contribui com: (0,998 - 1) *100 = -0,2%. Ou seja, para cada doação a mais de pessoa física, diminui em 0,2% a chance de estar entre os partidos grandes. Para número de candidatos o efeito é: (1,004 - 1)*100= 0,4%. Para o total de operações de doações ao partido fica: (1,001 – 1)*100 = 0,1%. Portanto, as contribuições positivas também são muito baixas, com 0,4% para número de candidatos e 0,1% para total de doações. Como vemos, as variáveis inseridas no modelo não são boas para explicar a diferença entre partido grande e partido pequeno nesse caso. Tanto é assim que as medidas de intervalo de confiança (últimas do modelo) estão passando por zero entre o mínimo e máximo. Assim, podemos dizer que o que determinou se um partido terminasse como grande ou pequeno na eleição para deputado federal em 2014 não foi nem o número de candidato, nem o volume doadores e menos ainda as doações de pessoas físicas. A regressão nos mostrou que a explicação está em outras variáveis. Por fim, o teste de colinearidade VIF ficou abaixo de 10,0 para as três variáveis, indicando inexistência de inflação artificial de efeitos por colinearidade.

Neste capítulo, discutimos três tipos de regressões: linear simples, linear múltipla e logística binária. Existem muitas outras fórmulas de regressão entre variáveis. Nos próximos capítulos, discutiremos algumas adaptações dos coeficientes de regressão para testes ou variáveis que não se enquadram nos pressupostos dos modelos canônicos de regressão.

7.6 Referências bibliográficas do Capítulo VII

Bunchaft, G., & Kellner, S. R. O. (1999). *Estatística sem mistérios*. Vol. II. Petrópolis: Editora Vozes.

Freijo, J. B. (2013). El paquete estadístico R. *Cuadernos metodológicos*, 48. Madrid: Centro de Investigaciones Sociológicas.

Gujarati, D. (2006). *Econometria Básica*. Rio de Janeiro: Editora Campus.

Menard, S. (1995). *Applied Logistic Regression Analysis*. London: Sage Publications.

Triola, M. F. (1999). *Introdução à Estatística*. Rio de Janeiro: Editora LTC.

7.7 Exercícios propostos do Capítulo VII

7.7.1 A partir do banco de dados BDC7V2_LEG, disponível em https://blogempublico.files.wordpress.com/2018/02/bdcap8v2_leg.xlsx, considere o que segue:

- Variável dependente: “votnom” total de votos nominais em candidatos do partido para deputado federal em 2014.
- Variável explicativa 1: “doapfis”, número de operações de doações de pessoas físicas ao partido.
- Variável explicativa 2: “Doar.tot”, total em R$ arrecadados pelo partido em 2014.
- Variável explicativa 3: “numcand”, total de candidatos apresentado pelo partido para deputado federal.
- Variável explicativa 4: “votleg”, total de votos de legenda obtidos pelo partido na eleição para deputado federal em 2014.

O objetivo é testar o ajustamento do modelo e os efeitos individuais das quatro variáveis independentes sobre o total de votos obtidos pelos partidos. Então:

a) Rode uma regressão linear múltipla e analise os principais coeficientes.

b) Rode um gráfico de resíduos e interprete as distribuições.

c) Rode o teste de heteroscedasticidade de *Breusch-Pagan* e o teste de colinearidade VIF. Analise os resultados. Se necessário, proponha alternativas para solucionar os problemas apresentados pelo modelo inicial.

7.7.2 Utilizando o mesmo banco de dados, rode uma regressão logit onde:

- Variável dependente: “partgr”, binária onde 1=partido grande e 0=partido pequeno.
- Variável explicativa 1: “Doart.tot”, total em R$ arrecadados pelo partido em 2014.
- Variável explicativa 2: “numcand”, total de candidatos a deputado federal apresentados pelo partido em 2014.

O objetivo é testar a contribuição dos recursos de campanha e do número de candidatos para a posição do partido nas categorias “grande” ou “pequeno” para a eleição de deputado federal de 2014. Analise:

a) As contribuições individuais por Bexp e odds ratio;

b) O VIF das variáveis explicativas.

c) Interprete os resultados para responder o que contribuiu mais para um partido ser grande na eleição de deputado federal em 2014?

Capítulo VIII

Análise de Trajetória (*path analysis*)

Quando as coisas acontecem em sequência, isso interfere em como elas acontecem (Tilley, 1993).

A primeira aplicação específica dos testes de regressão que veremos é a da análise de trajetórias ou análise de dependência (em inglês, *path analysis*). Nesta técnica, o objetivo é verificar como se dão as associações entre variáveis que estão separadas no tempo ou entre aquelas em que é possível identificar algum efeito conjunto que não pode ser aferido diretamente na equação da reta da regressão múltipla. Ou seja, a análise de trajetórias é uma técnica que se aplica quando se pretende fazer uma análise múltipla de efeitos, considerando o efeito de uma variável explicativa sobre a outra e não apenas os efeitos controlados de todas as explicativas sobre a dependente.

Para melhor compreender a funcionalidade da análise de trajetória, pense em uma regressão linear cuja variável dependente é nota obtida na prova final do semestre de uma turma de alunos e as variáveis explicativas são duas: a) nota obtida na primeira prova do semestre e b) tempo de estudo em horas por semana. É evidente que podemos considerar as duas variáveis (a) e (b) como explicativas do desempenho dos alunos na última prova semestral. A nota na primeira prova é preditora da nota na seguinte, assim como o tempo de estudo em horas semanais também ajuda a explicar o desempenho escolar dos alunos. Nos dois casos, espera-se uma associação positiva com intensidades

distintas. No entanto, perceba que tempo de estudo em horas semanais também tem influência sobre a nota na primeira prova dos alunos, logo, tempo de estudo é anterior à primeira nota. Em análise de trajetória, o tempo de estudo passa a ser uma variável explicativa da nota na segunda prova e moderadora do efeito da nota da primeira prova. Assim, a diferença neste tipo de análise é que a técnica torna possível quantificar o efeito direto do tempo de estudo sobre a nota na prova final, além do efeito indireto do tempo de estudo, aquele que é capturado pela nota na primeira prova. Como veremos neste capítulo, apesar de algumas limitações em sua aplicação (a principal delas é ser restrita a variáveis intervalares ou de razão), a análise de trajetórias oferece uma riqueza de detalhes nos resultados que a regressão linear clássica não é capaz de fornecer.

8.1 Princípios da análise de trajetória

A análise de trajetória (*path analysis*) ou análise de dependência é uma técnica preocupada em identificar as relações e não necessariamente as causas (Duncan, 1966). O coeficiente de trajetória, assim como a formalização do modelo, foi usado pela primeira vez em 1920, pelo geneticista Sewall Wright, quando a utilizou em suas pesquisas sobre a seleção natural. Duncan (1966) mostra que a análise de trajetória começa com a produção de um diagrama com caixas e flechas indicando as direções dos efeitos. Os valores parciais no diagrama são chamados de coeficientes de trajetória, dados pelo Beta padronizado de uma regressão linear. Também é possível encontrar modelos que usam como coeficientes de trajetória o coeficiente de correlação de Pearson entre duas variáveis. Porém, na ciência política esse uso é pouco indicado, pois não permite a identificação da direção da relação. A soma dos efeitos parciais (diretos e indiretos) indica os efeitos totais das variáveis explicativas e intervenientes sobre a dependente. Para Duncan (1966), não pode haver ambiguidade a respeito da ordem temporal das variáveis em um modelo de trajetória. As datas de ocorrências são o ponto de partida para o desenho de uma estrutura de efeitos diretos e indiretos, ou seja, para a organização do diagrama de caixas e flechas.

De acordo com Lipset e Rokkan (1967) o ponto de partida de toda análise de

trajetória é reconhecer que padrões específicos no tempo e suas sequências importam para a explicação dos fenômenos. Ainda que com o mesmo ponto de partida, duas variáveis podem apresentar efeitos distintos ao final, sendo que custos iniciais interferem nos efeitos finais e as diferenças nos efeitos de trajetória permitem distinguir o que é formativo daquilo que é efeito de reforço no processo (Lipset & Rokkan, 1967). Segundo esses autores, existem quatro fatos que caracterizam a dependência de trajetória como técnica de análise dos fenômenos políticos: i) múltiplo equilíbrio, que produz retornos crescentes a um grande número de resultados possíveis; ii) contingenciamento, pois eventos relativamente pequenos, quando ocorrem no momento adequado, podem ter grandes efeitos; iii) a importância do tempo e da sequência nos processos políticos; e iv) a inércia, como resultado dos retornos recebidos que leva a um equilíbrio e dificulta a mudança.

De forma complementar ao que apresentam Lipset e Rokkan (1967), Pierson (2000) reforça que a análise de trajetória exige respeito a alguns pressupostos básicos. O primeiro é que os padrões devem estar associados às sequências temporais e que essas sequências importam. O segundo é que ainda que contem com pontos de partida similares, as trajetórias podem levar a resultados distintos. Por fim, como consequência disso, uma vez estabelecida a trajetória, ela não pode ser revertida (Pierson, 2000).

Alwin e Hauser (1975) se preocupam em definir e diferenciar efeitos direitos, indiretos e totais em contextos de análise de trajetória. Segundo eles, uma associação entre duas variáveis sempre indica a relação direta entre elas. Porém, a associação – ou efeito – total de uma sobre a outra depende também de um termo de efeito indireto, que não é randômico. Caso fosse aleatório, não seria efeito indireto e sim termo de erro. Portanto, dado o viés do termo de efeito indireto, ele pode ser acrescido ao modelo explicativo e seus coeficientes calculados pelas técnicas de mínimos quadrados. O efeito total de uma variável sobre a outra é dado pela relação entre as duas, independentemente do mecanismo pelo qual as associações ocorrem. O efeito indireto é a parte do efeito total que é transmitido via mediação de outra variável, que intervém na associação entre as duas originais do modelo especificado. Já o efeito direto é aquela parte do efeito total que não foi transmitida pela variável interveniente ou moderadora (Alwin & Hauser, 1975). Em outras palavras, é o efeito que sobra se a mediadora fosse uma constante.

A análise de trajetória é uma técnica que parte dos modelos de regressão linear

múltipla para identificar os pesos individuais das explicações de cada variável independente sobre a dependente a partir da mediação de uma trajetória específica ou algumas outras. Ao realizar testes empíricos, ela permite uma aproximação entre o modelo teórico – o das representações gráficas nos diagramas – e as estatísticas empíricas. A técnica acrescenta à regressão o conceito de variável mediadora. Quer dizer, trata-se da utilização dos coeficientes de regressão clássica, porém adicionando a eles a ideia de mediação de efeitos. Por exemplo, os processos eleitorais estão entre os fenômenos sociais nos quais a sequência temporal é fundamental. Eventos prévios importam mais do que outros e diferentes sequências no modelo podem produzir resultados distintos, o que justifica a análise do desempenho partidário em eleições realizadas ao longo do tempo ao invés de explicações sincrônicas, como se a eleição fosse um evento discreto, quando na verdade é um contínuo no tempo (Cervi, 2016). Mas, antes de falar do efeito de mediação, é preciso diferenciar mediação de moderação, pois o efeito de moderação em uma regressão é distinto do da mediação. Os dois são importantes, pois acrescentam qualidade explicativa sobre os efeitos da variável explicativa sobre a dependente, porém distintos. O efeito de moderação é gerado pela inclusão de uma variável moderadora na regressão. Um moderador é um fator que influencia a relação dos demais no modelo. Já a variável mediadora é inserida em um teste de regressão para verificar a força da relação anterior das variáveis. Vejamos primeiro o efeito moderador.

Em uma regressão com modelo convencional a variável explicativa é representada pela letra X e a dependente pela letra Y. Aqui, a variável moderadora é representada pela letra Z. Seu papel é influenciar um modelo de regressão e essa influência ocorre quando uma relação entre X e Y varia como uma função de Z. Isso indica que existia um tipo de relação entre X e Y antes da inserção da moderadora e outro tipo após a entrada da variável Z. Quando Z tem efeito moderador sobre a relação anterior, significa que a correlação entre X e Y não é consistente, dada a distribuição de Z como representado em $[Z \rightarrow X \rightarrow Y]$. Em outras palavras, a relação entre X e Y varia em função de diferentes níveis de Z. Aplica-se o conceito de variáveis moderadoras em regressões apenas quando as variáveis são contínuas. Para saber se existe efeito de Z sobre a relação de X e Y, basta comparar os coeficientes r^2 do modelo anterior à moderação e do posterior à moderação. Se em um teste de comparação de médias os coeficientes mostrarem-se

estatisticamente distintos, isso significa que houve efeito de moderação pela variável Z.

Já a mediação é outro efeito possível de se encontrar na inclusão de uma terceira variável em uma relação anterior. A mediação é considerada um mecanismo de efeito, importante para quando se quer compreender como uma variável mediadora (M) intermedeia a relação entre X e Y. Quando X e Y estão correlacionadas, podemos usar um teste de regressão para predizer o valor de Y a partir de X, como discutido no capítulo anterior. Também é possível que X e Y estejam correlacionados devido à mediação da variável M. Então, a variável mediadora gera um efeito porque ela encontra-se entre a variável X e Y como representado em: $[X \rightarrow M \rightarrow Y]$. A questão a ser verificada aqui é se o coeficiente Beta continua sendo significativo após o efeito de mediação. Se o coeficiente (B) da relação original for reduzido após a inserção da variável mediadora é sinal de que houve um efeito de mediação. Se o Beta continuar o mesmo, significa que M não exerce efeito de mediação. Espera-se que todo mediador (M) tenha algum efeito sobre a relação entre X e Y. Se houver mudança, ainda que pequena, então se diz que existe uma mediação parcial. Se a mudança for completa, fala-se em mediação total (Edward & Lambert, 2007).

8.2 Componentes do modelo de análise de trajetória

A análise de trajetória (*path analysis*) é um tipo de teste aplicado principalmente para verificar o efeito de mediação ou moderação. Essa metodologia usa ilustrações que indicam os componentes de trajetória. As figuras que compõem as representações são as seguintes:

- Retângulos: indicam as variáveis observadas (Z, X, M, Y);

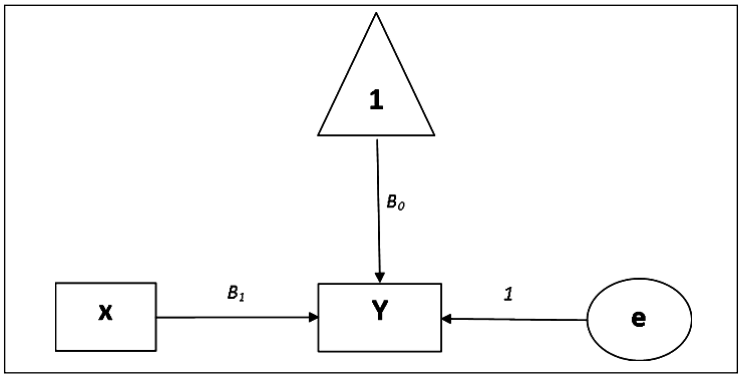
- Círculos: indicam as variáveis não observadas. Ou seja, os fatores de erro (e), aqueles que não podem ser medidos por serem aleatórios;

- Triângulos: indicam as constantes. Ou seja, as características que não variam no modelo e, portanto, não podem ser consideradas variáveis;

- Flechas: indicam as direções das associações. Elas só possuem direção quando se utiliza o beta padronizado no modelo.

Os elementos apresentados acima fazem parte dos diagramas de trajetória que devem anteceder todos os modelos e a produção de coeficientes, como descrito na representação 8.1 a seguir:

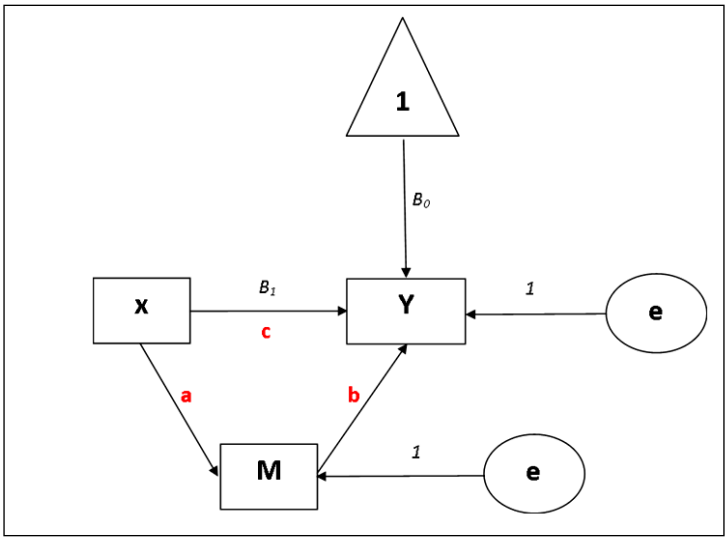
Figura 8.1. Representação básica da equação de regressão linear simples



Fonte: autor

O modelo acima é uma representação gráfica de variáveis relacionadas em uma regressão linear simples. Onde a variável dependente Y sofre três efeitos distintos: o efeito B_0 que é constante; um efeito B_1 que é o efeito da variável explicativa X sobre ela e um terceiro efeito que é o erro presente no modelo. Em termos matemáticos, a representação seria na fórmula tradicional para o cálculo da regressão: $Y = B_0 + B_1X + e$. Quando inserimos a variável mediadora moderadora, a representação gráfica passa a ser a representada na figura 8.2:

Figura 8.2. Representação do modelo de regressão com variável Mediadora



Fonte: autor adaptado de Edward e Lambert, 2007.

Agora, além do efeito direto de X sobre Y, há também um efeito mediado de X passando por M para chegar até Y. Em um modelo como o representado na figura 8.2, é possível medir o "tamanho" desse efeito e estabelecer se a variável M apresenta um efeito forte de mediação sobre X e Y ou não. Aqui, a direção da flecha aponta para a direção da influência. Nesse caso, X influencia Y diretamente e também através de M. Podemos representar cada efeito parcial substituindo os valores pelas letras na cor vermelha do modelo acima.

a: Trajetória de X para M;

b: Trajetória de M para Y;

c: Trajetória direta de X para Y (desconsiderando o efeito de M).

Note que se multiplicarmos a por b teremos o chamado coeficiente de trajetória indireta da X sobre Y passando por M. Ou seja, teremos o coeficiente parcial de trajetória que não é identificado diretamente na relação entre as variáveis X e Y.

Como já descrito aqui, o primeiro passo da análise de trajetória é a representação teórica dos efeitos das variáveis em um modelo que é composto pelos retângulos, círculos e setas em posições determinadas. A análise de trajetória mais comum é feita a partir da comparação entre diferentes modelos, alterando gradativamente o número de variáveis para comparar os resultados. Começa-se com o modelo máximo, que inclui todas as variáveis em que há expectativa de efeitos. Depois, são produzidos modelos "reduzidos", só com variáveis significativas, para comparar e identificar qual delas apresenta o modelo com maior capacidade explicativa.

Nesse tipo de análise de trajetória, o diagrama reúne as seguintes variáveis:

- Var. Independentes (exógenas) = que não têm causas explícitas sobre o fenômeno e/ou estão separadas no tempo. Representadas por X;

- Var. Intermediárias (endógenas) = são as imediatamente anteriores ao fenômeno e que se espera que apresentem efeitos explícitos sobre a dependente. Representadas por Z no caso de moderadora ou por M no caso de mediadora;

- Var. dependente = é a que representa o fenômeno que se pretende explicar. Representadas por Y.

Para produzir qualquer modelo de trajetórias o primeiro passo é posicionar as variáveis e as setas nas direções das explicações. Em seguida, calculam-se os coefi-

cientes parciais de regressão. Na figura 8.2, o efeito de X sobre M é representado pela letra (a) e pode ser substituída pelo coeficiente Beta padronizado da regressão linear simples de X como independente e M como dependente. O efeito de M sobre Y é representado pela letra (b) e o termo matemático dele no modelo é o coeficiente B padronizado da regressão linear simples de M como variável independente Y como dependente. O terceiro termo é o efeito direto de X sobre Y, representado pela letra (c) e com coeficiente Beta padronizado produzido a partir da regressão linear simples, com X como variável independente e Y como dependente. Assim, temos os três termos (coeficientes B padronizados) parciais dos efeitos diretos de X sobre M, de M sobre Y e de X sobre Y. Para obter os termos completos do modelo, basta multiplicar os termos parciais entre si, respeitando as direções das setas e depois somá-los. Um ponto crucial do modelo é posicionar corretamente as variáveis X e M, pois a inversão entre elas pode alterar os resultados finais. Em geral, a lógica determina qual o fator com efeito anterior, no caso, a variável X, e qual o efeito entre as duas variáveis originais, no caso a variável M. Em outros modelos, quando há uma diferença temporal de ocorrências fica mais evidente a diferença entre as variáveis explicativas e intermediárias. A explicativa sempre antecede temporalmente as intermediárias, que estão mais próximas no tempo da variável dependente. Já nos modelos de mediação acontece o contrário. A variável Z tende a aparecer antes no tempo que a variável X para explicar Y.

8.3 Aplicação do modelo de análise de trajetória

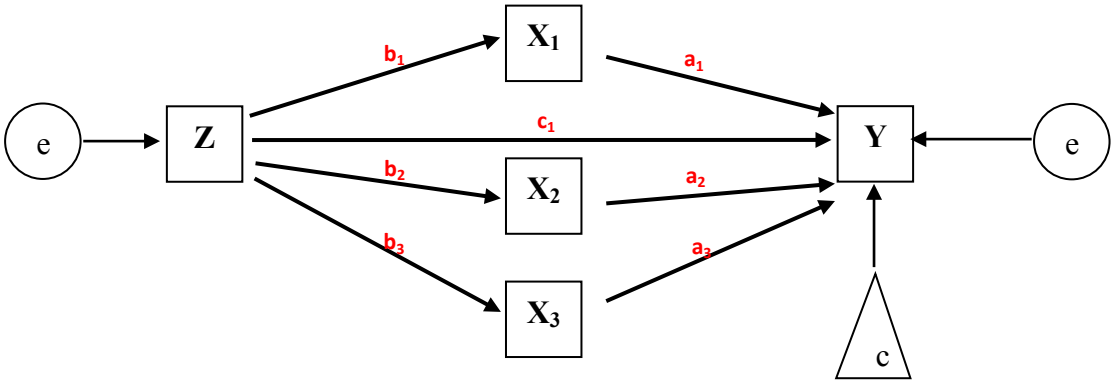
Toda análise de trajetória começa com a composição do diagrama de variáveis e setas, que servirá para identificar a antecedência temporal entre as variáveis e as relações entre efeitos diretos, indiretos e totais. No exemplo que usaremos aqui, vamos usar o banco de dados do capítulo 7 e agregar variáveis mediadoras e moderadoras. Então, para começar, relacionamos diferentes características das votações dos partidos para deputado federal para explicar o total de votos que cada partido obteve em 2014. O diagrama 8.1 a seguir sumariza as informações das variáveis por tipo e os cálculos para encontrar os efeitos totais. Temos como variável dependente, aquela que

queremos explicar as variações, o total de votos nominais obtidos pelos candidatos a deputado federal em 2014 por partido. Foram incluídas três variáveis explicativas no modelo de regressão múltipla: i) total de candidatos apresentados pelos partidos; ii) total de pessoas físicas que fizeram doações aos partidos ou aos candidatos a deputado federal em 2014; e iii) total de votos de legenda do partido. No capítulo 7, identificamos a partir da análise de regressão linear múltipla que essas variáveis apresentam efeitos estatisticamente significativos para explicar as variações de votos nominais. No entanto, no modelo de regressão linear não é possível inserir efeitos de fenômenos anteriores à eleição de 2014 como mediadores. No máximo conseguiríamos incluir variáveis anteriores como explicativas, ao lado das três independentes apresentadas acima.

Para exemplificar a análise de trajetória vamos incluir a variável "total de votos para deputado federal em 2010 por partido" como mediadora das relações entre as independentes e a dependente no modelo. O princípio é que o número de candidatos, o de doadores e o de votos de legenda em 2014 foram mediados pelo total de votos que o partido obteve na eleição anterior. Ou seja, o desempenho em 2010 ajuda a explicar a distribuição de candidatos, doadores e votos de legenda em 2014 e, por consequência, faz a mediação dos efeitos para o total de votos nominais obtidos pelos partidos em 2014. Perceba que o efeito a ser analisado aqui é o de mediação (M), pois o total de votos em 2010 é temporalmente anterior às variáveis explicativas e isso faz com que a mediação seja o efeito esperado entre os votos em 2010 e as variáveis independentes em 2014 para explicar as variações dos votos nominais aos partidos nesse ano. Se a trajetória a ser identificada ocorresse na mesma disputa, por exemplo, com inclusão da variável total de receitas declaradas pelos candidatos em R$ para explicar os votos nominais, o efeito seria de moderação (Z), pois os recursos de campanha teriam efeitos sobre o número de candidatos, número de votos de legenda e número de votos nominais na mesma disputa e ainda apresentaria efeito direto sobre o total de votos obtidos pelos partidos. Utilizaremos o modelo de Mediação no exercício proposto a este capítulo. Para o exemplo, vamos verificar o efeito mediador do total de votos a deputado federal em 2010 para os votos nominais em 2014 e todas as variáveis explicativas inseridas no modelo, conforme apresentado no diagrama 8.1 a seguir.

Diagrama 8.1. Aplicação do modelo de trajetórias para votos em 2014 para dep. federal

Var. Moderadora	Var. Explicativas	Var. Dependente
Z. Total de votos ao partido na eleição anterior.	X_1. Total de candidatos do partido. X_2. Total de doadores (pessoas físicas) ao partido. X_3. Total de votos de legenda ao partido.	Y. Votos nominais ao partido na eleição atual.



Var. Explicativa	Efeito Total = (Ef. Direto + (Ef. Indireto))
X_1. Total de candidatos do partido em 2014.	$Et_{x1} = a_1 + (b_1 * a_1)$
X_2. Total de doadores pessoas físicas em 2014.	$Et_{x2} = a_2 + (b_2 * a_2)$
X_3. Total de votos de legenda em 2014.	$Et_{x3} = a_3 + (b_3 * c_3)$
Z. Total de votos ao partido em 2010.	$Et_m = c_1 + ((b_1 * a_1) + (b_2 * a_2) + (b_3 * a_3))$

Y = Total de votos nominais obtidos pelos candidatos do partido para deputado federal em 2014
e = proporção da variação não explicada pelas variáveis independentes.
c = características que não variam, são as constantes. Ex.: número de vagas em disputa.

A hipótese a ser testada é a de que os votos nominais em 2014 para deputado federal estão relacionados diretamente ao total de candidatos apresentados pelo partido + total de doadores pessoa física do partido + total de votos de legenda obtidos pelo partido + total de votos em 2010. Mas, o efeito de cada uma das três primeiras variáveis explicativas sobre os votos nominais em 2014 também sofre a mediação do total de votos obtidos em 2010. Considera-se que o desempenho do partido em 2010 tem impacto sobre o número de candidatos, doadores e votos de legenda em 2014, que por sua vez ajudam a explicar as variações de votos nominais em 2014.

O coeficiente de trajetória é o coeficiente (beta padronizado da regressão linear) que indica o efeito direto e individual de uma variável sobre a outra. Quando o modelo tem mais de uma variável independente e mediadora, cada coeficiente é um coeficiente parcial dos efeitos de trajetória. Antes de calcular os coeficientes dos efeitos totais, a técnica de análise de trajetória exige o cumprimento de duas etapas prévias. A primeira é comparar todos os modelos de regressão possíveis a partir do modelo máximo especificado

no diagrama 8.1 para identificar qual apresenta maior variação explicada. Espera-se que o modelo com o maior número de variáveis seja o mais explicativo. Se não for, é preciso reconsiderar a especificação inicial, alterando – normalmente pela exclusão de algumas variáveis explicativas iniciais. A segunda etapa prévia é a geração dos modelos de regressão para termos os coeficientes (beta padronizado) parciais. No nosso caso, será um modelo para gerar os coeficientes a_n, outros para os coeficientes b_n, e um terceiro para o coeficiente c_1. Em seguida, basta calcular os efeitos totais a partir dos diretos e indiretos.

1ª ETAPA DA ANÁLISE – identificação do modelo com maior capacidade explicativa.

Aqui, rodaremos modelos de regressão com as variáveis explicativas e a moderadora para identificar as diferentes capacidades explicativas das variações de cada um deles. Os resultados usados para a definição do melhor modelo são o resíduo bruto do modelo, o erro bruto e a estatística F de cada um deles. No nosso exemplo, podemos propor quatro modelos de regressão, como seguem:

MOD1: *votnom2014 ~ cand2014+ doadores2014+ votoslegenda2014+ totalvotos2010.*

Onde “total de candidatos”, “total de doadores”, “total de votos de legenda em 2014” e o “total de votos em 2010” explicam os votos nominais em 2014 para deputado federal. É o modelo máximo e considera que quanto maior o número de candidatos, de doadores e de votos de legenda da eleição atual, além do total de votos na eleição anterior, maior tenderá a ser o número de votos nominais a candidatos dos partidos.

MOD2: *votoslegenda2014 ~doadores2014 +cand2014+totalvotos2010.*

Onde “total de doadores” e “total de candidatos” pelo partido em 2014 e o “total de votos em 2010” explicam os votos de legenda em 2014. Quanto mais votos na eleição anterior, mais candidatos e doadores na eleição atual, maior o número de votos de legenda no partido é a hipótese que sustenta o teste.

MOD3: *doadores2014 ~ cand2014+totalvotos2010*.

Onde "total de candidatos em 2014" e "total de votos em 2010" explicam o número de doações de pessoas físicas em 2014. Quanto mais votos na eleição anterior e mais candidatos na eleição atual, maior tenderá a ser o número de doadores (pessoas físicas).

MOD4: *cand2014 ~totalvotos2010*

Onde "total de candidatos em 2014" é explicado pelo "total de votos em 2010". Quanto mais votos na eleição anterior, maior a probabilidade de ter mais candidatos na eleição seguinte.

Usando o *RCommander* vamos rodar os quatro modelos de regressão, cujos resultados são apresentados na saída abaixo. Como estamos comparando duas eleições, a de 2010 e a de 2014, mantivemos no banco de dados para a análise apenas os partidos que apresentaram candidatos a deputado federal em ambas as disputas. Isso fez com que o número de casos ficasse em 27 partidos (N = 27). Para facilitar a leitura dos resultados, foram mantidos na saída apenas o r² e a estatística F de cada modelo, que são usadas na identificação da capacidade explicativa das variações de cada um deles.

```
LRcmdr> MOD1 <-lm(VOTNOM14~TOTCAND14+DOAPFIS14+VOTLEG14+TotVot10,data=TRAJET)
Multiple R-squared:  0.9104,     Adjusted R-squared:  0.8941
F-statistic: 55.88 on 4 and 22 DF,  p-value: 3.292e-11

Rcmdr> MOD2 <-lm(VOTLEG14~DOAPFIS14+ TOTCAND14+ TotVot10, data=TRAJET)
Multiple R-squared:  0.7952,     Adjusted R-squared:  0.7685
F-statistic: 29.76 on 3 and 23 DF,  p-value: 0.00000004285

Rcmdr>  MOD3 <- lm(DOAPFIS14 ~ TOTCAND14 + TotVot10, data=TRAJET)
Multiple R-squared:  0.6659,     Adjusted R-squared:  0.6381
F-statistic: 23.92 on 2 and 24 DF,  p-value: 0.000001933

Rcmdr>  MOD4 <- lm(TOTCAND14 ~ TotVot10, data=TRAJET)
Multiple R-squared:  0.4084,     Adjusted R-squared:  0.3847
F-statistic: 17.26 on 1 and 25 DF,  p-value: 0.0003331
```

Como se percebe nas saídas dos quatro modelos, o primeiro é o que apresenta o maior r² e também o maior coeficiente F, portanto, é o mais explicativo de todos os possíveis, conforme esperado inicialmente. Como todos os coeficientes F são estatisticamente significativos a 95% de intervalo de confiança, vale a pena realizar os cálculos

de erro e de resíduos para se certificar de qual modelo apresenta maior explicação de variações. O resíduo bruto é calculado para cada modelo da seguinte forma: (1 – r²). Quanto menor o resíduo bruto, maior a proporção de variação explicada no modelo.

MOD1: Resb = 1 – 0,910 = 0,089

MOD2: Resb = 1 – 0,795 = 0,205

MOD3: Resb = 1 – 0,665 = 0,335

MOD4: Resb = 1 – 0,408 = 0,592

O modelo 1 apresenta o menor resíduo bruto de todos eles, ou seja, é o que tem menos variância não explicada, indicando que ele oferece a melhor explicação entre as alternativas. Para calcular o erro bruto de variância, também chamado por coeficiente de trajetória, basta extrair a raiz quadrado do resíduo bruto de cada modelo. O erro bruto de variância ou coeficiente de trajetória é representado pela letra "e". Então, temos:

MOD1: $e = \sqrt{(1 - r^2)} = \sqrt{0{,}089} = 0{,}298$

MOD2: $e = \sqrt{(1 - r^2)} = \sqrt{0{,}205} = 0{,}452$

MOD3: $e = \sqrt{(1 - r^2)} = \sqrt{0{,}335} = 0{,}578$

MOD4: $e = \sqrt{(1 - r^2)} = \sqrt{0{,}592} = 0{,}769$

O menor erro bruto de variância continua sendo, também como esperado, o do modelo 1 (e=0,298), indicando que este modelo é o melhor especificado entre todos os possíveis. Assim, concluímos a etapa 1 da análise de trajetória com a identificação de que o modelo proposto no quadro 8.1 é o mais adequado para explicar as variações dos votos por partido para deputado federal em 2014. Agora passamos à segunda etapa. Não há necessidade de ser redundante e passar pelos três cálculos. Apenas um indicador (r², resíduo bruto ou erro bruto de variância) já é suficiente para identificar a correta especificação do modelo.

2ª ETAPA DA ANÁLISE – Identificação dos coeficientes individuais

Para identificar os coeficientes individuais é preciso ter os valores de Beta padronizado (que são padronizações das estimativas de Beta em unidades de desvio padrão) para cada variável. Nesse caso, precisamos rodar dois modelos para obter os coeficientes padronizados. O modelo A nos fornecerá os coeficientes de Beta padroni-

zado das variáveis X_n e Z para Y. Os modelos B_n nos fornecerão os coeficientes da variável Z para cada uma das variáveis X, como descrito no diagrama 8.1. Então, teríamos:

RegA: VoNom14 ~ totcand14 + doafis14 + votleg14+totvot10.

Essa regressão nos fornecerá um coeficiente beta padronizado para cada uma das quatro variáveis explicativas iniciais. Além dela, teremos:

RegB1: totcand14 ~ totvot10.

RegB2: doafis14 ~ totvot10.

RegB3: votleg14 ~ totvot10.

Cada uma dessas regressões acima nos fornecerá o beta padronizado da regressão do total de votos em 2010 para cada uma das variáveis independentes do modelo.

ETAPA FINAL DA ANÁLISE – cálculo dos coeficientes parciais

Assim que forem encontrados todos os coeficientes padronizados, basta substituí-los no modelo do diagrama 8.1 para, em seguida, fazer as multiplicações e somas necessárias para os efeitos diretos e indiretos entre as variáveis. Uma limitação do *RCommander* é que o teste de regressão linear múltipla dele não fornece automaticamente o coeficiente beta padronizado. Ele dá apenas o Beta, chamado de Estimativa de cada variável. A fórmula para calcular o Bpad é:

$$Bp = B\ x\ \frac{Sx}{Sy}$$

Onde Sx e Sy são as variâncias de X e Y.

Uma das maneiras de se obter o Beta padronizado no *RStudio* é baixando o pacote ("*QuantPsyc*") e usando o comando [>*lm.beta* (regressão linear)]. Este comando fornecerá os valores de beta padronizado para todas as variáveis independentes inseridas no modelo de regressão linear original (O anexo 8.6 apresenta o passo a passo da geração do beta padronizado usando o pacote "*QuantPsyc*" no *RStudio*). No nosso exemplo, os coeficientes das regressões estão na saída de resultados abaixo – para simplificar, serão reproduzidos aqui apenas os valores de beta padronizado de cada modelo na sequência apresentada na segunda etapa da análise.

```
lm.beta(REGa)VOTNOM14
  DOAPFIS14    TOTCAND14     VOTLEG14     TotVot10
 -0.20225223   0.09004066   0.66685530   0.43144848

lm.beta(REGb1)TOTCAND14
 TotVot10
0.6390587

lm.beta(REGb2) DOAPFSI14
 TotVot10
0.7954425

lm.beta(REGb3)VOTLEG14
 TotVot10
0.8465887
```

Agora que já temos todos os coeficientes parciais de trajetória (Beta padronizados) podemos identificar os efeitos parciais, diretos e indiretos, e os efeitos totais no modelo de trajetória. Para chegar aos efeitos totais sobre os votos nominais em 2014 é preciso somar os efeitos diretos e os efeitos indiretos.

O efeito direto de totcand14 sobre votnom14 é de +0,090;

O efeito direto de doapfis14 sobre votnom14 é de -0,202;

O efeito direto de votleg14 sobre votnom14 é de +0,666;

O efeito direto de totvot10 sobre votnom14 é de +0,431.

Esses são os quatro efeitos diretos do modelo de trajetória sobre os votos nominais em 2014. O que mais explica individualmente as variações de votos nominais em 2014 é o número de votos de legenda em 2014 (+0,666), seguida do total de votos em 2010 (+0,431). O número de candidatos tem uma explicação baixa (+0,090). É curioso perceber que apresentar mais candidatos não significa ter mais votos nominais para deputado federal. Há um coeficiente negativo, de -0,202, para total de doador pessoa física. O sinal negativo mostra uma relação inversa entre as duas variáveis. O coeficiente de número de doadores é maior que o de total de candidatos para explicar as variações de votos nominais nos partidos.

O próximo passo é considerar os efeitos indiretos da variável moderadora (totvot10) sobre a variável dependente, ou seja, o impacto que ela tem sobre as variáveis explicativas do modelo de trajetória.

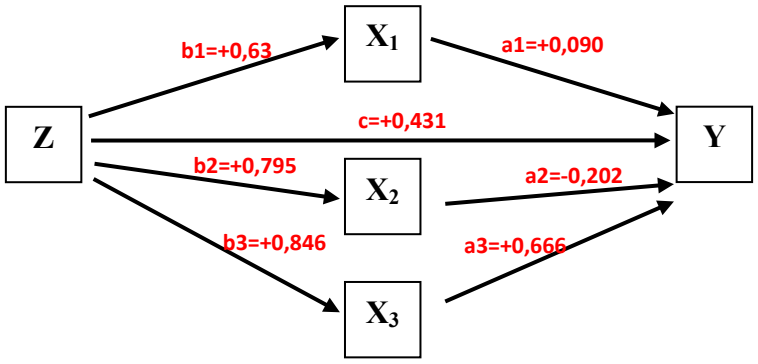
O efeito indireto de totvot10 via totcand14 é: 0,639 x 0,090 = +0,057;

O efeito indireto de totvot10 via doapfis14 é: 0,795 x (-0,202) = -0,160;

O efeito indireto de totvot10 via votleg14 é: 0,846 x 0,666 = +0,563.

Agora que já temos todos os coeficientes parciais, podemos substituir os valores no diagrama 8.1 para proceder aos cálculos dos coeficientes de trajetória, como segue no diagrama 8.2 abaixo. No mesmo diagrama foram inseridos os cálculos para obtenção dos efeitos totais de cada variável explicativa.

Diagrama 8.2. Resultados do modelo de trajetórias para votos em 2014 para dep. federal



Var. Explicativa	Efeito Total = (Ef. Direto + (Ef. Indireto))
X_1. Total de candidatos do partido em 2014.	$Et_{x1} = a_1 + (b_1 * a_1)$ = +0,090 + 0,057 = **+0,147**
X_2. Total de doadores pessoas físicas em 2014.	$Et_{x2} = a_2 + (b_2 * a_2)$ = -0,202 + (-0,160) = **-0,362**
X_3. Total de votos de legenda em 2014.	$Et_{x3} = a_3 + (b_3 * a_3)$ = +0,666 + 0,563 = **+1,229**
Z. Total de votos ao partido em 2010.	$Et_m = c_1 + ((b_1 * a_1) + (b_2 * a_2) + (b_3 * a_3))$ = 0,431 + (0,057 + (-0,160) + 0,563) = +0,413 + 0,478 = **+0,891**

Y = Total de votos obtidos pelo partido para deputado federal em 2014

A partir dos efeitos totais apresentados no diagrama 8.2, e considerando o modelo de trajetória montado, é possível afirmar em primeiro lugar que o principal efeito sobre os votos nominais em 2014, considerando a "memória eleitoral de 2010", foi o total de votos de legenda em 2014 – o que normalmente é desconsiderado nas análises de desempenho eleitoral partidário. Em seguida, apareceu o total de votos em 2010, usado aqui como a memória eleitoral do partido. Isso significa que, no cômputo, votos nominais para deputado federal em 2014 foi mais importante para o partido ter votos de legenda do que o montante de votos que ele havia obtido na eleição de 2010. Este resultado já tinha sido obtido no modelo de regressão linear total, aquele que considera todas as variáveis, inclusive na mesma ordem de intensidades de efeitos e de sinais dos coeficientes. Isso aconteceu porque a variável moderadora mediadora (total de votos em 2010) apresentou efeitos parciais muito próximos entre si para as três variáveis in-

dependentes (ver os coeficientes b1, b2 e b3 no diagrama 8.2). Isso mostra que o efeito de moderação mediação para o modelo foi baixo.

A mesma importância deve ser dada aos coeficientes baixos, pois a trajetória mostra que eles têm menor impacto sobre os votos nominais. É o caso da variável total de candidatos, indicando que lançar muitos candidatos explica pouco a variação de votos nominais entre os partidos. Assim como o sinal negativo da relação entre número de doadores (pessoa física) e votos nominais. O coeficiente indica que partidos que tendem a ter mais operações de doações de pessoa física são os que tendem a apresentar menor votação nominal ao final do processo. Em resumo, o principal achado aqui foi o efeito dos votos de legenda sobre os votos nominais ser maior que as demais variáveis explicativas, mesmo quando moderado pelo total de votos no partido na eleição anterior.

O objetivo principal de uma análise de trajetória é especificar melhor o modelo explicativo a partir das relações entre as variáveis. No caso do exemplo, fica claro que é mais forte a associação entre os votos nominais e votos de legenda no mesmo ano do que entre votos nominais em uma eleição e o total de votos alcançado pelo partido na eleição anterior. No caso dessa técnica de análise, o acréscimo de informação se dá pela possibilidade de relacionar efeitos parciais (diretos e indiretos) entre variáveis separadas no tempo. Para isso, é importante que o pesquisador tenha muito cuidado na hora de especificar o modelo e na construção do diagrama de variáveis e setas indicativas das relações diretas e indiretas da trajetória.

8.4 Referências bibliográficas do Capítulo VIII

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in *path analysis*. *American Sociological Review*, 4, 37-47.

Cervi, E. U. (2016). PT e PSDB em eleições subnacionais (1994 a 2014). *Anais do X Encontro da Associação Brasileira de Ciência Política* (ABCP), Belo Horizonte,

MG, Brasil, 2016.

Duncan, O. D. (1966). Path analysis: sociological examples. *The American Journal of Sociology*, 72(1), 1-16.

Edwards, J. R., & Lambert, L. S. (2007). Methods for Integrating Moderating and Mediation: A General Analytical Framework Using Moderated *Path analysis*. *Psychological Methods*, 12(1), 1-22.

Lipset, S. M., & Rokkan, S. (Eds.). (1967). *Party systems and voter alignments: Cross-national perspectives*. Toronto: The Free Press.

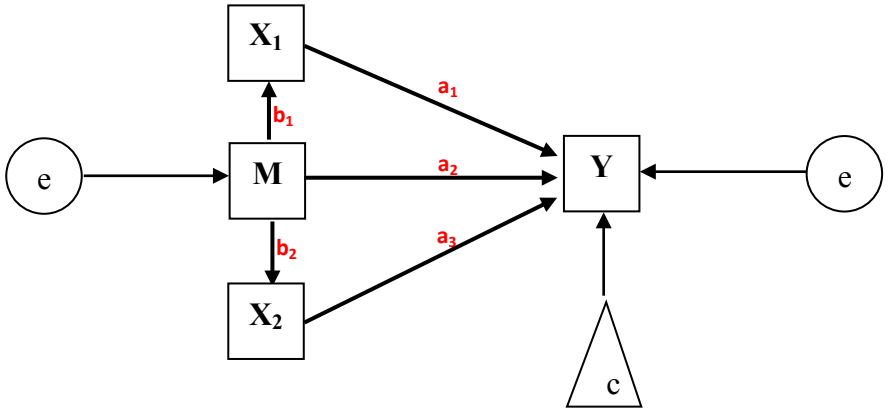
Pierson, P. (2000). Increasing Returns, Path Dependence, and the Study of Politics. *The American Political Science Review*, 49(2), 251-267.

Tilley, J. A. (1993). Valuing American Options in a Path Simulation Model. *Investiment Section Monograph*, 1, 55-67.

8.5 Exercícios propostos do Capítulo VIII

8.5.1 Baixe o banco de dados BDCAP9V2_TRAJ disponível em: https://blogempublico.files.wordpress.com/2018/02/bdcap9v2_traj.xlsx. Calcule os coeficientes totais do modelo de trajetórias apesentado no diagrama abaixo. Considere como variável dependente o total de votos obtidos pelos partidos em 2014 na eleição para deputado federal. Considere como variáveis independentes o total de doadores aos partidos (que é a soma do número de operações de doações) e o total de recursos de campanha (R$) declarados pelos candidatos a deputado em 2014. Considere como variável moderadora o número de candidatos apresentados pelos partidos. Espera-se que o total de candidatos em uma eleição tenha efeito sobre o total de doadores aos partidos e sobre o total de recursos recebidos para as campanhas em R$. Além disso, espera-se que as três variáveis tenham efeitos sobre o total de votos obtidos pelo partido. Siga as etapas da montagem do modelo de trajetórias, calcule os coeficientes e interprete os resultados.

Var. Moderadora	**Var. Explicativas**	**Var. Dependente**
M. Total de candidatos a deputado federal.	X_1. Total de recursos R$ declarados. X_2. Total de doadores aos candidatos e partidos.	Y. Total de votos ao partido em 2014.



Var. Explicativa	**Efeito Total = (Ef. Direto + (Ef. Indireto))**
X_1. Total de recursos R$ declarados por candidatos em 2014.	$Et_{x1} = a_1 + (b_1 * a_1)$
X_2. Total de doadores a candidatos e partidos em 2014.	$Et_{x3} = a_3 + (b_2 * a_3)$
M. Total de candidatos a deputado federal em 2014.	$Et_m = a_2 + ((b_1 * a_1) + (b_2 * a_3))$

Y. Total de votos obtidos pelos partidos para deputado federal em 2014.
e. proporção da variação não explicada pelas variáveis independentes.
c. características que não variam, são as constantes. Ex.: número de vagas em disputa.

ANEXO DO CAPÍTULO VIII

ANEXO 8.1 – PASSO A PASSO PARA PRODUÇÃO DA ESTATÍSTICA BETA PADRONIZADO DA REGRESSÃO LINEAR NO *RSTUDIO*:

```
# INSTALAR O PACOTE QuantPsyc
> install.packages("QuantPsyc")


# CARREGAR O PACOTE "QuantPsyc"
> library(QuantPsyc)


# CARREGAR O BANCO DE DADOS
> attach(BDCAP9V2_TRAJ)


# RODAR A REGRESSÃO LINEAR ENTRE AS VARIÁVEIS
> mod1<- lm (VOTNOM14~TOTCAND14+DOAPFIS14+VOTLEG14+TotVot10)


# SOLICITAR AS ESTATÍSTICAS BÁSICAS DA REGRESSÃO (NÃO OBRIGATÓRIO)
> summary(mod1)

Call:lm(formula = VOTNOM14 ~ TOTCAND14 + DOAPFIS14 + VOTLEG14 +
TotVot10)

Residuals:
      Min        1Q    Median        3Q       Max
-10804548  -3153185    510881   2002200  24013278

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.149e+06  3.198e+06  -1.297 0.207973
TOTCAND14    2.314e+04  2.286e+04   1.012 0.322491
DOAPFIS14   -1.735e+03  1.055e+03  -1.645 0.114144
VOTLEG14     1.438e+01  3.040e+00   4.729 0.000102 ***
TotVot10     2.381e+00  7.002e-01   3.400 0.002569 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7009000 on 22 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9104,   Adjusted R-squared:  0.8941
F-statistic: 55.88 on 4 and 22 DF,  p-value: 3.292e-11


# SOLICITAR O BETA PADRONIZADO DAS VARIÁVEIS EXPLICATIVAS
> lm.beta(mod1)
  TOTCAND14   DOAPFIS14    VOTLEG14    TotVot10
 0.09004066 -0.20225223  0.66685530  0.43144848
```

Capítulo IX

Análise Geográfica

Tudo está relacionado com tudo, mas coisas mais próximas estão mais relacionadas do que coisas mais distantes. Waldo Tobler (primeira lei da geografia).

A análise geográfica é, antes de tudo, a interpretação de cores, pontos e linhas sobre representações espaciais. Os mapas são como gráficos por representarem visualmente o que há de mais importante, porém, nunca representam a totalidade dos fenômenos geográficos. Ainda que seja usada como técnica de análise há mais de 100 anos[1], a difusão de bases de dados geográficas a partir dos anos 1970 permitiu o desenvolvimento de análises exploratórias espaciais específicas que são usadas para estudos de fenômenos políticos com instrumentos de pesquisa mais elaborados (Anselin & Rey, 1991). O passo seguinte foi a apresentação de técnicas regressivas específicas para análises geográficas, aplicadas principalmente a descrições de fenômenos eleitorais. O objetivo do capítulo é introduzir o leitor às técnicas básicas de análise espacial descritiva e rudimentos de análise de regressão espacial. Para uma abordagem detalhada da técnica, ver Rodrigues-Silveira (2013).

A característica mais importante dessas técnicas é que elas solucionam o proble-

[1] Considera-se a primeira obra com uso de bases geográficas para análise política o livro de André Siegrifed, *Tableau Politic de la France*, publicado em 1913, no qual o autor relaciona os padrões de voto com as diferenças do tipo de solo do norte e sul da França (Terron, 2012). Para uma discussão resumida das relações entre princípios teóricos e técnicas de análise geográfica aplicada a fenômenos políticos, ver Zavala (2012)

ma criado pela quebra do pressuposto de independência entre os resíduos presentes nos testes tradicionais de associação e de medição da dependência entre variáveis. Como veremos em detalhes neste capítulo, a análise geográfica permite identificar dependência geográfica dos resíduos entre as variações dos fenômenos políticos e controlar seus efeitos, tornando os resultados mais precisos (Anselin & Rey, 1991), dado que um pressuposto básico das técnicas regressivas clássicas é a independência entre os resíduos. O capítulo apresenta um breve histórico da técnica, seguido da apresentação de *softwares* específicos para esse tipo de análise e que tenham código aberto. Como exemplo para a potencialidade de análises será introduzido o Geoda, *software* específico para análise de *clusters* espaciais. A partir dele serão apresentadas estatísticas descritivas univariadas, seguidas de análises clássicas de associação entre variáveis. Técnicas de análise espacial I de Moran, de *clusters* locais "Lisa" e, por fim, a aplicação de técnicas regressivas com dependência geográfica por defasagem e por erro espacial.

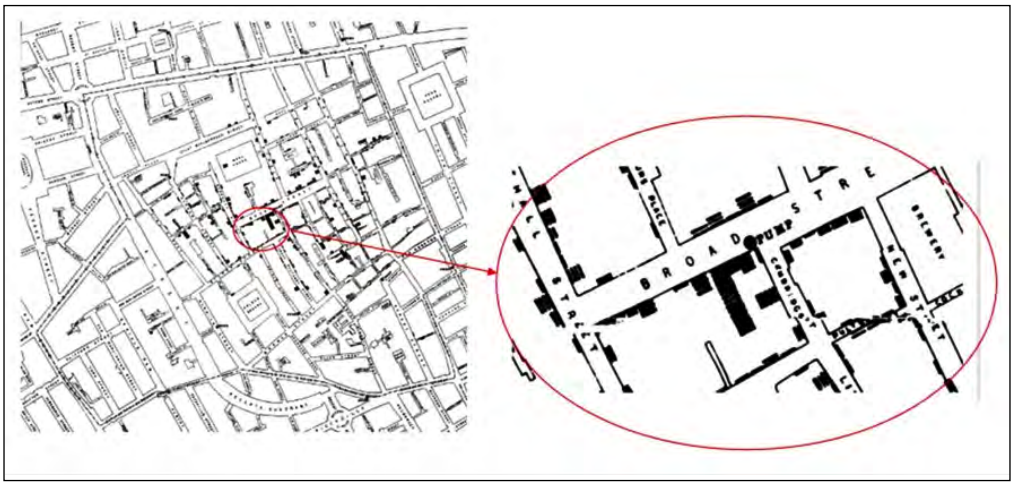
9.1 Princípios e objetivos da Análise Geográfica

A técnica de análise geográfica, também chamada de espacial, é outra variação dos princípios básicos de uma regressão. No entanto, aqui, o que se busca inicialmente é identificar possíveis relações das ocorrências entre unidades espaciais distintas. Ou seja, na análise geográfica, o espaço importa. Aqui é preciso fazer uma distinção fundamental para não incorrermos no risco da falácia ecológica. Quando afirmamos que o espaço importa, não podemos transferir automaticamente os resultados para os indivíduos ou instituições que estão presentes nesse espaço. A unidade de análise aqui não são pessoas ou grupos, mas sim a unidade espacial e tudo que está dentro dela. Assim, concluir que determinado município ou País apresenta dada característica que o distingue de seus vizinhos não nos permite afirmar nada sobre os cidadãos ou as instituições individualmente que compõem essas unidades de análise.

A análise geográfica deve ser entendida como uma extensão dos tipos de análises tradicionais, que buscam conexões entre pessoas ou instituições. Ela identifica como o espaço influencia as relações entre os atores políticos e as instituições a partir

de dados agregados e não individuais. Pode-se, com isso, descobrir determinados padrões ou anomalias espaciais. Mais de meio século antes da publicação do trabalho de Siegrifed, que tinha por objetivo apresentar os princípios da análise geográfica, a técnica já tinha sido usada na área de saúde pública, ainda que sem a preocupação de descrever seus princípios. Foi durante um surto de cólera, em Londres, em 1854 (Zavala, 2012). O médico John Snow, percebendo que as abordagens individuais de controle da doença não funcionavam, decidiu considerar o espaço como variável explicativa, como indica a reprodução do mapa abaixo.

Ilustração 9.1. Reprodução e detalhe do mapa de Snow com indicação de casos de cólera em Londres



Fonte: domínio público

Explorando o mapa da cidade de Londres, ele marcou os pontos dos óbitos por cólera, identificando inicialmente a densidade espacial da doença. Em seguida, identificou os locais de fornecimento de água da cidade e traçou os polígonos de Thissen para identificar as fronteiras naturais de um mapa a partir da distância entre pares de pontos. O pressuposto é que ocorrências dentro de um polígono sempre estarão mais próximas entre si do que em relação às de fora dele. Com isso, foi possível identificar quais fontes de água estavam mais próximas das maiores densidades de mortos por cólera e, ao impedir que a população tivesse acesso a esses poços, houve maior controle da doença. A partir de então, em mais de um século e meio, as técnicas de análise geográfica têm avançado em diferentes áreas do conhecimento.

Em um trabalho de 1983, Johnston faz uma revisão de afirmações feitas até então sobre a variabilidade individual nas análises de geografia eleitoral, ou seja, trata da clássica questão da falácia ecológica. Ele distingue os estudos eleitorais em três escalas principais: i) macro, envolvendo resultados nacionais, por exemplo, distribuição das cadeiras de um parlamento nacional por partidos; ii) resultados subnacionais, análise de desempenho eleitoral por distritos ou municípios ou outro grupamento subnacional qualquer, tendo como principal objetivo identificar quem ganha e quem perde em cada área; e iii) micro, análises individuais de tomada de decisão de voto, nesse caso, perguntando a cada eleitor sobre os critérios adotados para escolha do partido ou candidato.

De acordo com Johnston (1983), a análise geográfica eleitoral encontra-se no grupo (ii) de estudos, pois seu principal interesse é analisar diferentes padrões de votação delimitados por unidades espaciais. Ele defende ser possível considerar a existência de uma variação individual das decisões com a estabilidade dos resultados eleitorais ao mesmo tempo. Para isso, aplica o conceito de **oscilação uniforme**, demonstrando que a variação individual e a estabilidade geográfica são compatíveis. Isso porque a maior parte das oscilações individuais tende a ser anulada mutuamente, o que mantém a estabilidade na escala macro, no caso, subnacional. Para ele, o resultado de todos os fluxos decisórios, inclusive de se abster, é o volume líquido de mudança partidária e, além de ser menor que as mudanças individuais, isso tende a ser consistente ao longo do tempo. Por este princípio, Johnston (1983) afirma que a manutenção dos resultados no nível macro é compatível com as oscilações no nível micro. Em um segundo momento, o pesquisador defende que existe um ganho para a geografia eleitoral quando se analisam os resultados espaciais ao longo do tempo, pois isso permite não apenas a identificação de padrões de voto, como também de manutenções e mudanças deles ao longo do tempo, em diferentes eleições.

Em estudo posterior, Johnston *et al.* (2001) questionam a ausência de variáveis contextuais nas análises geográficas de apoio a partidos políticos na Grã-Bretanha e uma tendência de sobreposição de explicações socioeconômicas. Para eles, o espaço é uma variável mais importante para explicar a influência sobre as escolhas eleitorais do que as próprias características sociais. Demonstram existir uma variação maior nas preferências partidárias entre tipos de áreas do que entre diferentes tipos de pessoas. Assim, defendem que, por exemplo, o apoio ao partido conservador é maior em áreas rurais e em

distritos de transições do rural para o urbano, sendo mais baixo em áreas urbanas com concentração de jovens e trabalhadores braçais. Antes, Books e Prysby (1991) propuseram uma teoria dos efeitos contextuais que identifica quatro processos possíveis para análise da tomada de decisão eleitoral: a) pela observação pessoal, onde o voto de um eleitor é influenciado pela situação do seu entorno, como por exemplo, a saúde da economia local; b) a partir da interação, via comunicações interpessoais entre pessoas próximas entre si que acaba influenciando a decisão de voto; c) por interação em espaços organizacionais formais, como locais de trabalho, igreja, sindicato, escola e outras organizações que possuem espaços para discussão de questões políticas; d) a partir dos meios de comunicação social que fornecem conteúdos focados em temas políticos relevantes sobre eventos próximos aos eleitores. Em todos os quatro casos há uma influência do espaço nas interações e contexto no qual se encontram os eleitores para a tomada de decisão de voto. Os processos de tomada de decisão influenciados por conversações são o que Miller (1977) definiu como “pessoas que falam juntas, votam juntas”. Trata-se do **efeito de vizinhança**. Assim, o mapa de distribuição espacial de votos tende a ser mais explicativo que a distribuição de votos por características socioeconômicas.

Potter & Olivella (2015) estudaram estratégias eleitorais usando a unidade geográfica para análise e concluíram que a proximidade entre distritos permite aos partidos melhor uso e ganho em escala das estruturas de campanha. O que parece ser óbvio intuitivamente teve testado o seu mecanismo pelas técnicas de regressão espacial. Os autores mostram que a proximidade entre distritos eleitorais amplia o efeito das estratégias ideológicas dos partidos e isso é o que estimula um partido a buscar apoio no distrito vizinho ao que ele tem domínio. A esse efeito eles dão o nome de efeito aditivo da geografia na tomada de decisão política.

Os estudos usando geografia eleitoral são antigos no Brasil, mas nas últimas décadas a técnica tem se difundido de forma acelerada. O objetivo aqui não é fazer uma revisão da literatura nacional na área, apenas apontar trabalhos que usaram diferentes abordagens e unidades espaciais. Soares (1973) analisou a desigualdade eleitoral no Brasil entre as décadas de 1950 e 1960 em função das regras que estabeleciam diferenças regionais na representação política. Mais recentemente, Terron e Soares (2010) analisam as mudanças nas bases eleitorais de Lula e do PT por unidade espacial em

sucessivas eleições nacionais. Braga e Rodrigues-Silveira (2011) também analisam as mudanças no padrão geográfico de votos para presidente da república em sucessivas eleições brasileiras. Para uma discussão conceitual e revisão do uso da técnica de análise geográfica eleitoral no Brasil, ver Terron (2012). Alkmin (2014) utiliza as mesmas técnicas de geografia do voto, porém, tomando como unidade de análise os bairros do Rio de Janeiro para analisar o desempenho dos candidatos a governador do Estado entre 1982 e 2010. No caso brasileiro há uma tendência de se reunir a análise geográfica com as séries temporais em estudos sobre desempenho eleitoral.

9.2 Bases de dados e *softwares* para Análise Geográfica

Na prática, o ponto de partida da análise geográfica é plotar informações em um mapa que permitam identificar semelhanças e diferenças entre as unidades espaciais. Para tanto, os *softwares* de análise espacial usam três tipos de arquivos básicos, que são chamados de shapefile. O primeiro deles é o arquivo shape, com extensão ".shp", no qual está o índice que permite indicar os limites dos polígonos, que representam as unidades espaciais. O segundo tem a extensão ".shx" e apresenta o ponto em que se encontra o centro de cada polígono. O terceiro é um arquivo de atributos, ou seja com os dados, na extensão ".dbf", que está relacionado a cada ponto dentro dos polígonos. Assim, uma informação qualquer em um arquivo ".dbf", pode ser localizada em determinado ponto do espaço ".shx", que por sua vez está contido dentro dos limites de um polígono ".shp". A partir daí é possível identificar semelhanças e diferenças entre unidades espaciais próximas ou distantes entre si e fazer regressões espaciais para descrições ou predições de comportamentos geográficos (nunca individuais).

Existe uma variedade de *softwares* com código aberto para análise geográfica, não sendo necessário recorrer a programas com código-fonte fechado. Aqui, vou citar apenas alguns, entre os mais conhecidos e os desenvolvidos no Brasil que são úteis para iniciantes em análise geográfica. Em todos os casos há manuais de instalação e introdução nos links apresentados a seguir. O mais conhecido é o **QGIS**, antigo Quantum Gis. Ele é bastante útil por ser georreferenciado em diferentes sistemas operacionais

e permite a análise por sucessivas camadas tanto em formato raster, quanto vetorial. Pode ser acessado em: https://www.qgis.org/pt_BR/site/. Há também o **Philcarto**, desenvolvido por Philippe Waniez, com código totalmente aberto, porém, não trabalha com georreferenciamento, restringindo-se às análise cartomáticas. Acesso por: http://philcarto.free.fr/. No Brasil, o Instituto Nacional de Pesquisas Espaciais (INPE) desenvolveu o **Terraview**, um visualizador de dados geográficos, que trabalha com dados vetoriais – polígonos, linhas e pontos - e matriciais – imagens e grades - georreferenciados. Ele é totalmente em português e apresenta plugins para edição e formatação de documentos e relatórios. Link em: http://www.dpi.inpe.br/terraview_previous/index.php/. Por sua vez, o Instituto de Pesquisas Econômicas Aplicadas (IPEA) desenvolveu o **IpeaGeo**, também com código aberto, com funcionalidades de georreferenciamento para análises com dados socioeconômicos espaciais no Brasil. Pode ser acessado por: http://www.ipea.gov.br/ipeageo/. Por fim, o **Geoda**, o *software* que será usado neste capítulo para introdução à análise geográfica. Desenvolvido pelo Centro de Dados Espaciais da Universidade de Chicago, também possui código aberto. Permite visualização de *clusters* a partir de mapas coropléticos e a realização de testes básicos de estatísticas descritivas, até diferentes tipos de regressões espaciais. Link em: https://spatial.uchicago.edu/geoda.

A primeira etapa da análise espacial descritiva é a atribuição de padrões em uma superfície. Isso se faz com o uso de mapas coropléticos, que distingue cores e tons para indicar diferenças proporcionais de quantidades em cada unidade espacial. Esses mapas são a forma mais comum de representação descritiva dos comportamentos geográficos. Eles vão desde a diferenciação em uma variável binária (presença/ausência) por duas cores, até as proporções de presença de uma variável contínua por dezenas de tons e cores diferentes. Nos mapas coropléticos cada unidade espacial (unidade de análise) é diferenciada das demais em função da presença ou da intensidade de determinada característica. Portanto, seu principal elemento de distribuição é a cor. O esquema das cores pode ser quantitativo ou em escala, quando cada cor representa uma quantidade distinta das demais. Pode ser qualitativo, quando as cores indicam características distintas, sem que exista nenhum tipo de hierarquia. E pode ser ainda divergente, quando existe um ponto médio ou central a partir do qual as cores indicam distribuições crescentes em direções opostas. Para uma descrição detalhada e justificativa de que cores usar em mapas, sugiro

acessar a página ColorBrewer em http://colorbrewer2.org/.

Além do tipo de escala, o que determina o número de tonalidades e cores em um mapa é a forma/técnica para estabelecer os limites entre eles. A descrição dessas técnicas será feita a seguir, já com o uso do Geoda.

9.3 Mapas coropléticos no Geoda

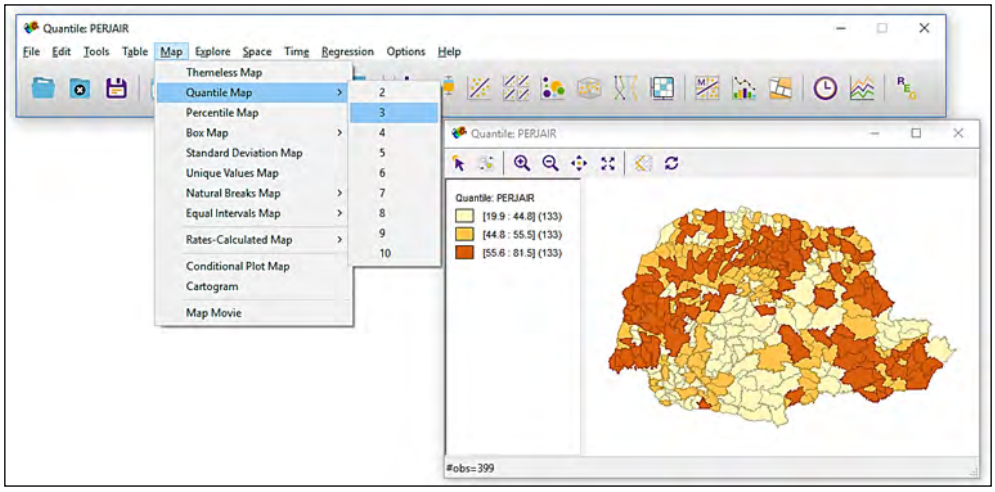
Devido aos objetivos do capítulo, não trataremos aqui da instalação e funcionalidades gerais do Geoda. Para isso, deve-se acessar a aba de tutoriais no portal do *software*. Como os procedimentos são simples e o programa é de fácil compreensão, vamos direto à descrição dos principais mapas gerados por ele. Para tanto serão usados dois manuais do Geoda de Luc Anselin: o manual revisado do usuário do Geoda (Anselin, 2003) e o manual que se encontra no site do *software*, link em: http://geodacenter.github.io/workbook/3a_mapping/lab3a.html. Como exemplo dos mapas, será utilizado um banco de dados com resultados eleitorais para municípios do Estado do Paraná. Este é o mesmo shape para os exercícios do final do capítulo, onde está o link para acessar a pasta com os arquivos.

O arquivo de dados tem como unidade de análise os 399 municípios do Paraná e contém informações sobre características socioeconômicas deles, posição geográfica e resultados eleitorais quantitativos como percentuais de votos a presidente, governador, senador e deputado federal em 2018, além do partido do prefeito eleito em 2012 e 2016 por cidade. Uma característica importante para a análise visual dos mapas coropléticos é que as variáveis quantitativas apresentem alguma normalização. Por exemplo, no caso de votos, indica-se a utilização de percentuais e não valores absolutos, pois em termos de valores absolutos, municípios com populações maiores sempre apresentarão maior intensidade de votos do que municípios com baixas populações. Por exemplo, qualquer candidato nanico tenderá a ter mais votos absolutos no município de São Paulo (com maior população do país) do que no município de Araguainha (MT), que possui uma das menores populações do Brasil. No entanto, quando normalizamos as votações em percentuais, é possível comparar diretamente o desempenho de um candidato na

cidade de São Paulo e Araguainha, pois não estamos mais tratando do total de votos obtidos, mas da proporção para o candidato em cada uma das unidades de análise. Por este motivo, todas as variáveis eleitorais quantitativas usadas aqui são referentes aos percentuais de votos obtidos pelo partido ou candidato no município.

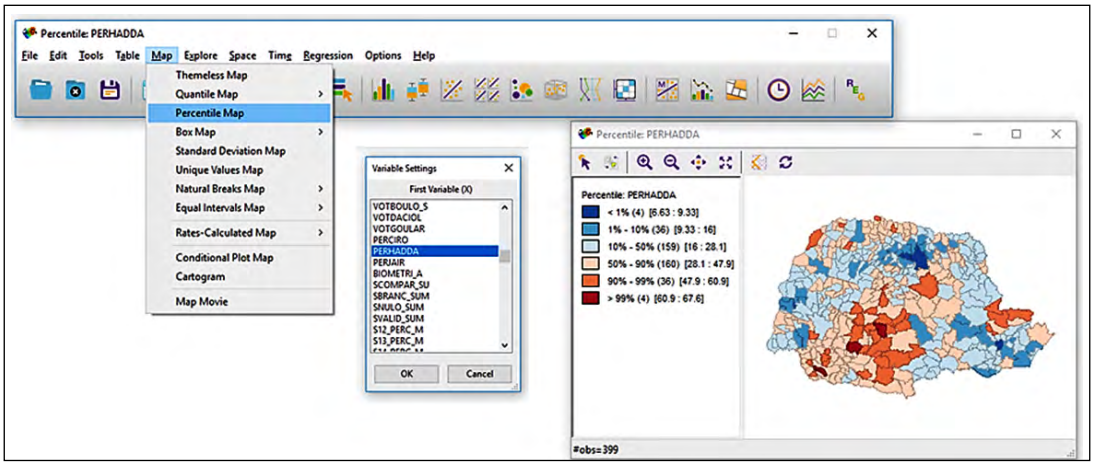
9.3.1 Mapa Quantil

O tipo mais comum de mapa coroplético é o quantil. Nele, são representadas as diferenças de intensidades a partir de grupos formados por um mesmo número de observações. No exemplo, foi solicitada a formação de três grupos. Com isso, o mapa produzido divide o total de municípios em três grupos de igual tamanho, com 133 municípios cada um. Aqui, o que varia é o intervalo de intensidades dentro de cada grupo, como é possível perceber na legenda ao lado do mapa. No caso, trata-se do percentual de votos em Bolsonaro (PSL) no primeiro turno de 2018. No primeiro grupo, com tom menos intenso, estão os 133 municípios em que o candidato obteve menos votos, variando de 19,9% a 44,8% (variação de mais de 25 pontos percentuais). No segundo grupo, com tom intermediário, encontram-se os municípios cujo percentual de votos variou de 44,8% a 55,5% (variação de pouco mais de 10 pontos percentuais). E o terceiro grupo, tom mais intenso, apresentou variação de 55,6% a 81,5% (de 27 pontos percentuais). Como se percebe, o mapa quantil é indicado para quando se busca homogeneidade na distribuição do número de observações por categoria e não homogeneidade dentro das categorias.



9.3.2 Mapa Percentil

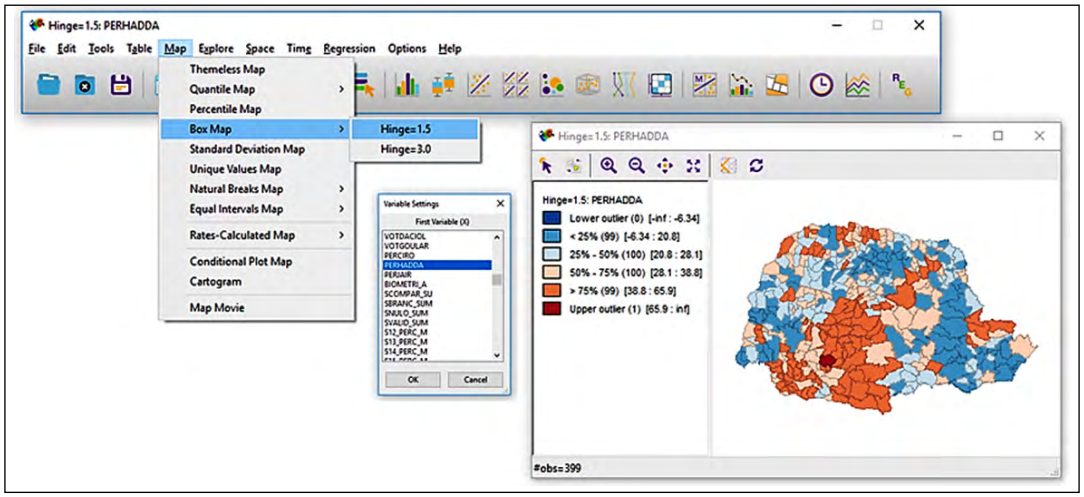
O mapa de percentil é indicado para identificação de casos extremos (*outliers*). Ele registra a distribuição dos casos em distintas faixas de percentuais. Mostra os valores contidos em 1% mais baixo dos casos (<1%), depois na faixa entre 1% e 10%, em seguida entre 10% e 50%, depois entre 50% e 90%, entre 90% e 99% e os *outliers* superiores, referentes a 1% superior (>99%). Além das diferenças de tons, são usadas duas cores. O azul, uma cor fria, para as unidades que ficam abaixo de 50% - quanto maior a intensidade de azul, mais baixa é a posição da unidade. A cor laranja para os casos acima de 50% - quanto maior a intensidade do laranja, mais próximo do extremo superior está o caso. No exemplo está a distribuição percentil da votação de Fernando Haddad (PT) nos municípios do Paraná no primeiro turno de 2018. No extremo inferior, 1% dos municípios, que representa um total de (4) na legenda, a votação dele variou entre 8,6% e 9,3%. Nos 10% inferiores (36 municípios), os percentuais ficaram entre 9,3% e 16%. Assim por diante, até o *outlier* superior, que são os 4 municípios com maiores percentuais de votos, com votação entre 60,9% e 67,6%. Nesse tipo de mapa o número de categorias é fixo em seis e as distribuições percentuais delas, também. Ele serve para indicar as distâncias entre os *outliers* e o ponto médio.



9.3.3 Box Map

É outro tipo de mapa para identificar a dispersão e valores extremos. Ele é apresentado em duas possibilidades (Hinge: 1.5 e Hinge: 3.0). A distribuição dos casos é em

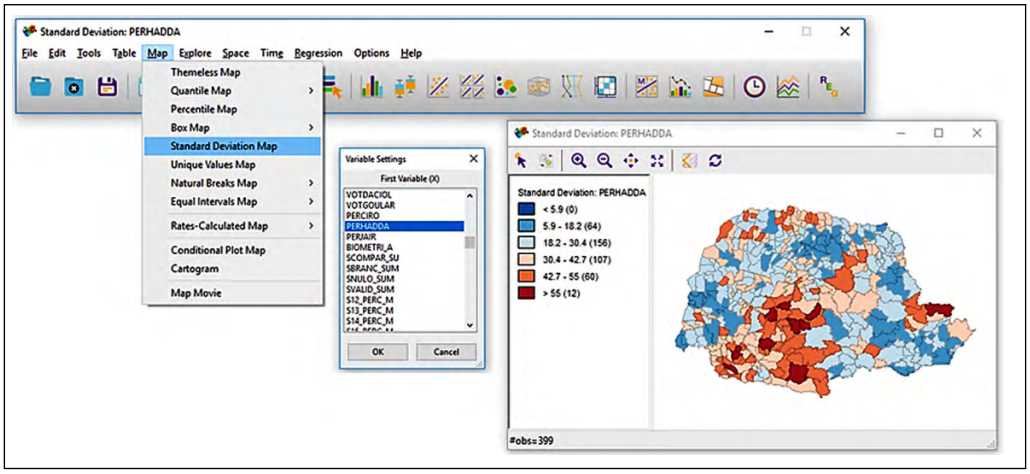
quartis, que em função da existência de *outliers* se transformam em seis categorias. Os limites entre as categorias são calculados a partir da mediana. Em Hinge 1.5, cada categoria está a 1.5 desvio padrão da mediana e os casos extremos estão na terceira categoria acima e abaixo da mediana. Com Hinge 3.0, os limites são calculados a partir de três desvios da mediana. Como se vê, com Hinge 3.0 há uma dispersão maior dos casos. Para comparar as diferenças de resultados, usaremos a mesma variável do mapa de percentil aqui. A distribuição dos percentuais de votos de Haddad, com Hinge 1.5 mantém a média de 28,1% de votos por município, mas indica que a 1,5 desvio padrão da média encontram-se 100 municípios acima e abaixo (25% dos casos para cada grupo). Não há *outlier* inferior segundo esse método e existe apenas um *outlier* superior, com 65,9% de votos.



9.3.4 Mapa de Desvio Padrão

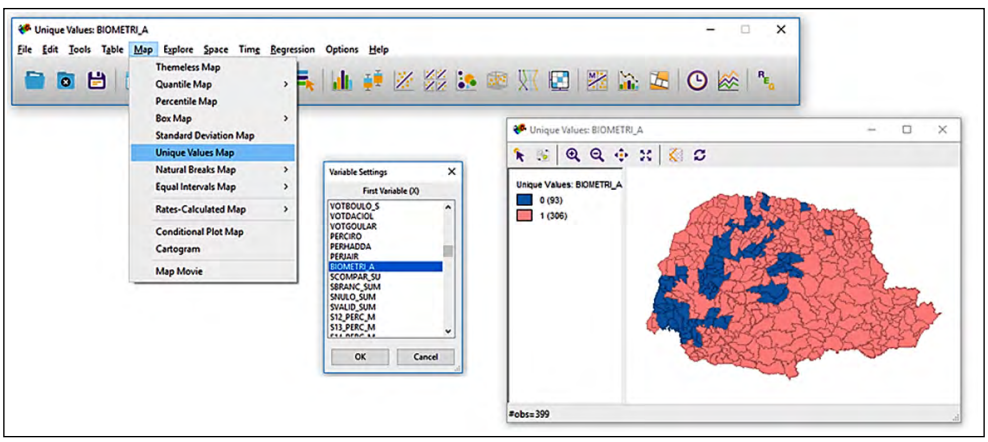
Outro mapa para identificar heterogeneidades entre as unidades de análise. Nesse caso, cada categoria inclui todas as unidades que se encontram a um desvio padrão da anterior. Ele começa do centro, a mediana, e divide as categorias por desvio padrão. Unidades abaixo da mediana recebem a cor azul. Acima da mediana são marcadas pela cor laranja. Quanto mais intenso o tom, mais distante do centro da distribuição. Aqui, o número de categorias depende da distância dos casos em relação ao centro da distribuição. Usando o gráfico para a mesma variável de distribuição de votos em Haddad, percebe-se que a média de 30,4% de votos apresenta 156 municípios com um desvio padrão para baixo, entre 18,2% e 30,4% e 107 municípios com um desvio para cima, variando de 30,4%

e 42,7% de votos. Não há nenhum município com três desvios a menos da mediana e existem 12 municípios acima de três desvios da mediana (> 55%).



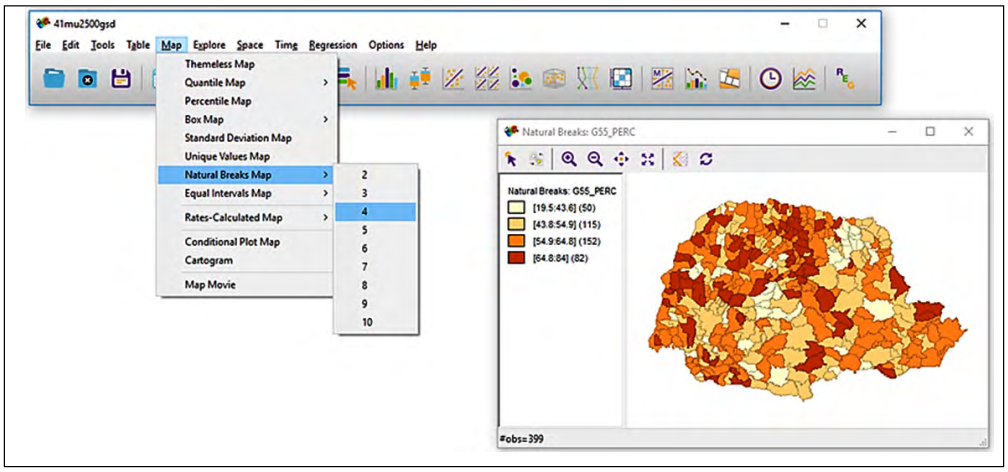
9.3.5 Mapa de Valores Únicos

Mapa para os casos de variáveis que não são quantitativas e no qual cada código diferente é indicado com uma cor. Serve para variáveis binárias, categóricas nominais ou ordinais. Não se recomenda mais que sete categorias para serem representadas no mapa. O exemplo é se em 2018 o município já tinha pelo menos iniciado o recadastramento biométrico ou se ainda era totalmente no sistema anterior de registro de eleitores, uma variável binária. No banco de dados, o código zero (0) é para municípios sem biometria e o código um (1) é para aqueles que tem biometria parcial ou total. No Paraná, em 2018 existiam 93 municípios sem biometria e outros 306 com biometria total ou parcial. Além disso, o mapa permite identificar que ainda não se fez recadastramento biométrico em municípios localizados em regiões específicas do sudoeste e do centro-norte do estado.



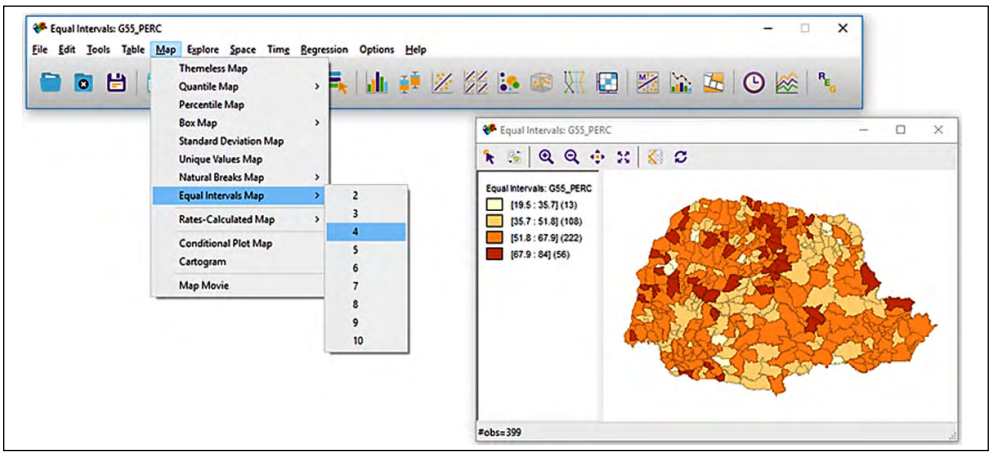
9.3.6 Mapa de quebras naturais

Uma opção para visualização de distribuições de variáveis quantitativas é o mapa de quebras naturais. De acordo com essa técnica, os limites dos grupos de uma variável quantitativa são definidos por um algoritmo que procura maximizar a homogeneidade interna de cada categoria. Ou seja, uma vez definido o número de grupos, a amplitude de valores de cada grupo vai variar em função do algoritmo proposto por Jenks (1977) para tornar os casos mais homogêneos possível dentro de cada categoria. No exemplo, é usada a distribuição de votos por município para o candidato a governador eleito em 2018, Ratinho Junior (PSD). A variável no banco de dados é a "G55_perc". Foi solicitado o agrupamento em quatro categorias. Como o algoritmo favorece a homogeneidade, há uma tendência de concentração de maior número de casos nas categorias centrais. No caso, a segunda categoria concentra 115 municípios com votação variando entre 43,8% e 54,9%. A terceira categoria apresenta 152 municípios onde as votações variam de 54,9% a 64,8%. Já a primeira e a última categorias têm menor número de municípios, pois encontram-se nos extremos. Note, ainda, que existe uma variação grande de amplitude entre as categorias. Ela vai de 23 pontos percentuais na primeira categoria (19,5% a 43,6%), passando a apenas 11 pontos percentuais na segunda categoria (de 43,8% a 54,9%). Esse tipo de mapa é indicado para quando não se busca intervalos com a mesma amplitude.



9.3.7 Mapa com intervalos iguais

Aqui, segue-se o mesmo princípio do mapa anterior. Uma variável quantitativa é agrupada em conjuntos de valores próximos entre si. O pesquisador define o número de categorias que quer produzir no mapa. A diferença é que a amplitude de cada categoria é sempre a mesma. O que varia é o número de unidades de análise nelas. Usando o mesmo exemplo da votação de Ratinho Junior, para fins de comparação, o mapa com intervalos iguais para quatro categorias produziu intervalos com amplitude de 15 pontos percentuais cada um, variando de um mínimo de 19,5% até o máximo de 84% de votos por município. Aqui é possível perceber que a terceira categoria concentrou o maior número de municípios (222), com percentuais entre 51,8% e 67,9%. Além disso, em comparação ao método de quebras naturais, há menor número de municípios nos extremos. Apenas 13 no primeiro grupo e 56 no quarto. Mapa com intervalos iguais é indicado para quando se quer comparar grupos com mesma amplitude, porém, com número de casos distintos.



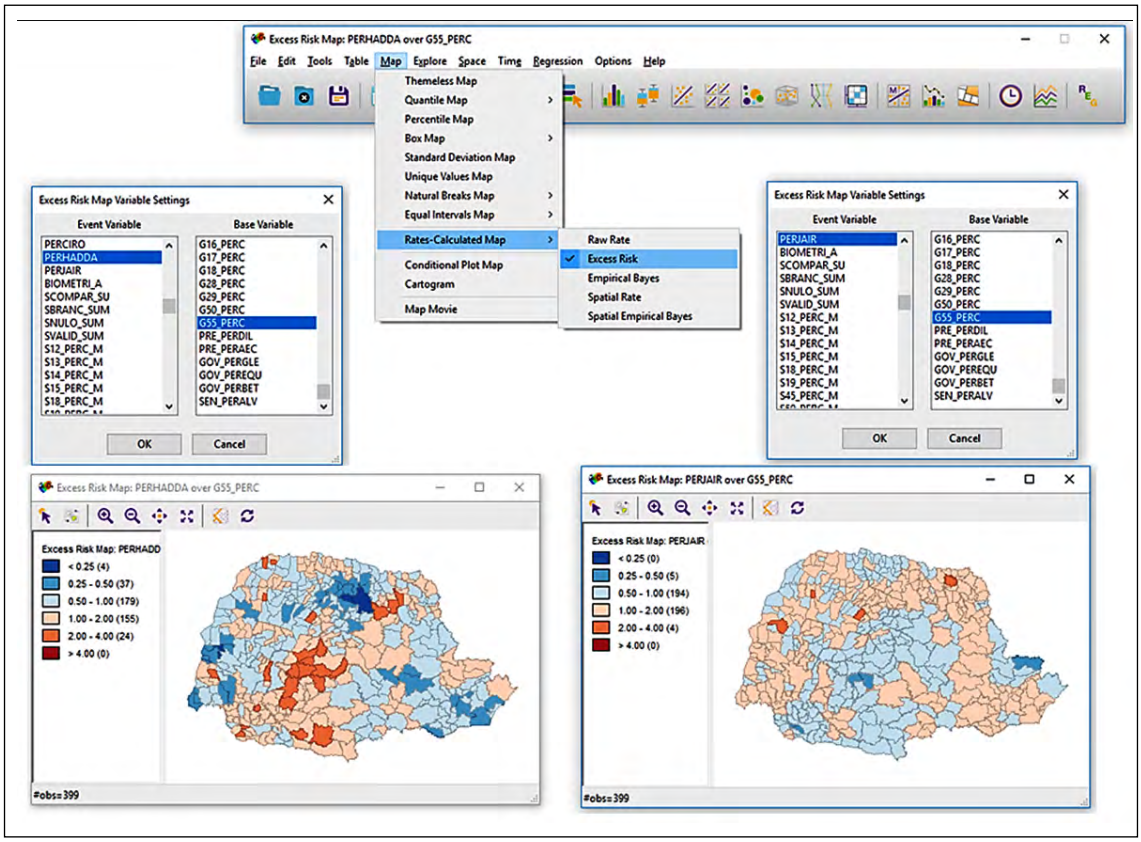
Até este ponto foram apresentados mapas coropléticos para análise descritiva de uma única variável no espaço. O Geoda também oferece algumas alternativas de mapas que consideram a relação entre duas variáveis no espaço. A seguir, será apresentada apenas uma delas, a que apresenta maior utilidade para análises eleitorais ou de políticas públicas

9.3.8 Mapa de razão de chance (*Excess Risk*)

Aqui, além das unidades espaciais, são utilizadas duas variáveis para calcular a razão de chance de ocorrência de uma delas em função da distribuição de valores da outra. Também é chamado de risco relativo ou excesso de risco. Para calcular, considera-se o valor médio de uma variável base para todas as unidades de análise. Sobre esse valor é calculado quanto que deveria existir da segunda variável, chamada de evento. Assim, se o valor real da variável evento ficar abaixo do esperado, ela é indicada com a cor azul. Se ficar acima, com laranja. As diferenças de tons mostram quanto o valor observado da variável evento ficou distante da ocorrência esperada. A principal utilidade do mapa de razão de chance em estudos eleitorais é para indicar quando que votos de dois candidatos na mesma eleição estão casados ou votos do mesmo candidato em duas eleições distintas coincidem ou não. No primeiro caso, a variável base é a votação de um candidato regional, enquanto a variável evento é a de um candidato a cargo nacional. Assim, o objetivo é identificar como se distribuiu a votação do candidato nacional, dada a votação do candidato regional. No segundo caso, a variável base é a primeira votação, ou eleição anterior. O evento é a eleição mais recente. O resultado indicará em que unidades espaciais houve diferenças para mais ou para menos na votação atual do candidato em relação à eleição anterior. As unidades espaciais marcadas na cor azul apresentaram resultado abaixo do esperado (< 1) para a variável evento, dada a distribuição da variável base. As unidades em laranja apresentaram resultado acima do esperado (>1), dada a distribuição da variável base. Convenciona-se considerar o primeiro intervalo abaixo e acima muito próximos do esperado.

A análise se dá no número de unidades duas ou mais categorias acima ou abaixo de um (1), que vão de 0,25 até 4 vezes o valor esperado. No exemplo abaixo usamos como variável base a votação do candidato a governador Ratinho Junior (PSD) e como variáveis evento as votações de Bolsonaro (PSL) e de Haddad (PT). Quanto mais unidades próximas a uma razão de chance de ocorrência (1), mais próximas foram os pares de votações nos municípios. O caminho para o mapa de razão de chance é o mesmo que os mapas anteriores, passando por Maps > Rates-Calculated Maps > *Excess Risk*. A partir daí abre-se uma caixa para selecionar a variável base e a evento. A votação de Ratinho Junior para governador (G55_perc) é a base para os dois casos. No mapa da esquerda o evento é a votação de Haddad e o da direita é a votação de Bolsonaro, no Paraná. Perceba que no caso de Bolsonaro, 390 dos 399 município apresentaram razão de chance em torno

de 1 (entre 0,5 e 2,0), o que indica um desempenho muito próximo do esperado, dada a distribuição de votos de Ratinho. Já no caso de Haddad os casos com razão de chance distantes de um crescem. São 24 municípios entre 2,0 e 4,0 vezes mais votos para Haddad que o esperado, dada a votação de Ratinho Jr. E em 37 municípios a votação de Haddad ficou abaixo do esperado. Isso mostra que os desempenhos de Ratinho Jr. e Bolsonaro foram “casados” nos municípios paranaenses, enquanto as de Ratinho Jr. e Haddad apresentaram maior independência. Além disso, nota-se um padrão regional nas distribuições de votos desses dois últimos. Haddad teve melhor desempenho que o esperado, dada a votação de Ratinho, principalmente em municípios da região sudoeste do estado. Se você olhar o desempenho de Haddad (mapa percentil, box map e mapa de desvio padrão acima) perceberá que essa foi a região de melhor desempenho eleitoral do candidato do PT. E se olhar o exemplo do mapa de quebras naturais, perceberá que foi a região que concentrou municípios com desempenhos baixos de Ratinho Jr. O que o mapa de razão de chance faz é unir essas duas informações, mantendo os municípios como unidade espacial de análise.



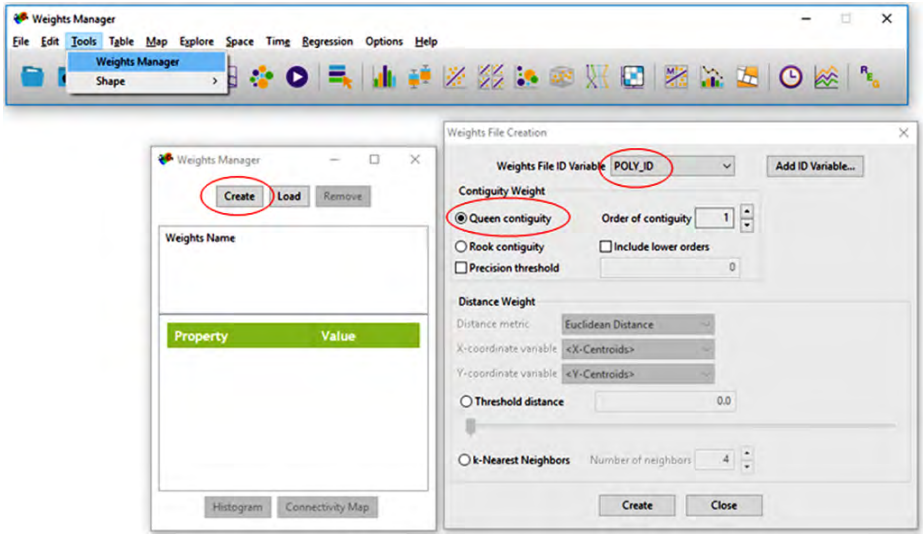
Neste tópico apresentamos as técnicas mais comuns de visualização de dados no Geoda, a partir de mapas coropléticos de intensidade e presença ou ausência de

valores. No próximo tópico avançaremos para análises com estatísticas e indicadores de associação geográfica, começando pelo coeficiente I de Moran, o Lisa para *clusters* locais e as regressões por lag e diferença de vizinhança.

9.4 Estatísticas básicas em Análises Geográficas Descritivas

Existem diferentes técnicas para realização da chamada Análise Exploratória de Dados Espaciais (AEDE), que são utilizadas para descrever e visualizar distribuições de proporções em unidades espaciais, verificar a existência de *outliers* e de *clusters* espaciais. Essas técnicas também são usadas na definição de modelos preditivos de desempenho no espaço. Um ponto importante para esse tipo de análise é a necessidade de definição de uma matriz de pesos espaciais, representado pela letra W. O objetivo dessa matriz é capturar os efeitos de contiguidade e vizinhança dos dados. O princípio é que duas unidades espaciais contíguas apresentam um peso maior do que duas unidades que não compartilham nenhum limite do polígono. Ou seja, unidades mais próximas têm peso maior que unidades mais distantes e quem identifica as proximidades é a matriz W. Existem diferentes formas de estabelecimento de matriz de pesos. A mais comum é a matriz Queen de vizinhança (1), que significa que todas as unidades espaciais que apresentem algum tipo de contiguidade recebe código 1 para peso de vizinhança.

No Geoda, o caminho para criar uma matriz de vizinhança é Tools > Weights Manager > Create > Add ID Variable > Poly_ID > Queen contiguity > Order of contiguity (1)



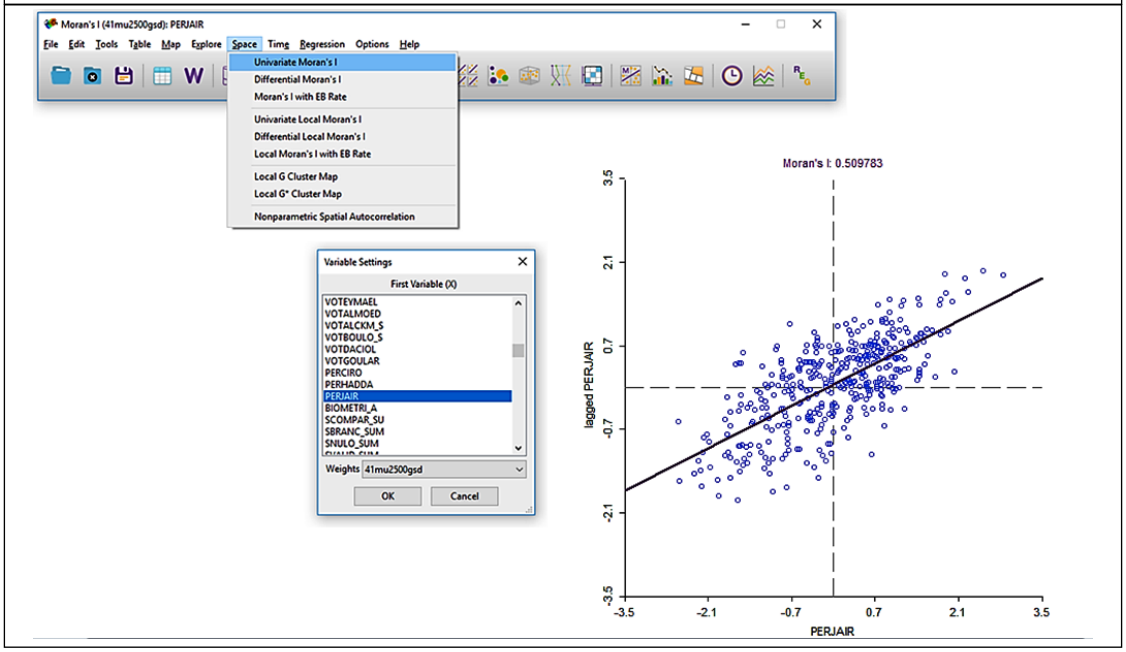
A caixa de diálogos para criação da variável peso espacial no Geoda apresenta como sugestão o nome “Poly_ID” para essa variável. Em seguida, basta escolher a opção “Queen” e ordem de contiguidade (1). Ao clicar em “Create” o programa acrescentará a variável “Poly_ID” com o peso por proximidade no arquivo “dbf” e criará um arquivo com extensão “.gal” na pasta dos arquivos para ser usado como peso nos testes exploratórios.

9.4.1 Autocorrelação espacial global com coeficiente I de Moran

O princípio da autocorrelação é que a proporção de determinada característica em uma unidade espacial tenderá a apresentar uma contiguidade nas unidades vizinhas. Ou seja, baixa proporção em uma unidade, deve indicar também baixa proporção em seus vizinhos. O mesmo se aplica às altas proporções. A ideia é que existe uma espécie de “transbordamento” de determinada característica entre os vizinhos no espaço. Quanto maior o coeficiente de I de Moran, maior a autocorrelação espacial global da variável analisada. Assim como outros coeficientes, o I de Moran varia do teórico 0,000, quando a autocorrelação espacial é nula e vai até 1,000, para autocorrelações espaciais perfeitas.

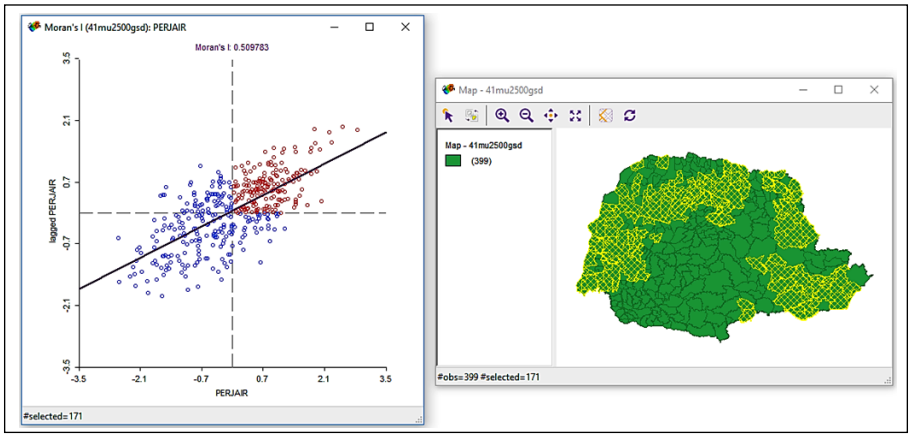
A representação da autocorrelação espacial não é feita em mapas, mas sim por gráficos de dispersão, como qualquer outro teste de correlação, onde cada ponto dentro do gráfico representa uma unidade espacial. O eixo X representa os valores identificado na unidade e o eixo Y mostra a média dos valores identificada na vizinhança da unidade, por isso é representada por “lag”. Ou seja, um retorno em relação à unidade de análise. Também é plotada uma reta de correlação entre as unidades espaciais e suas vizinhanças. Quanto mais inclinada for a reta, mais associação espacial entre os valores analisados. O princípio de interpretação do gráfico de dispersão I de Moran é o mesmo de qualquer teste de correlação já visto em capítulos anteriores. Como exemplo, usaremos o teste de I de Moran para medir a autocorrelação espacial global dos votos de Bolsonaro em 2018 nos municípios do Paraná.

No Geoda, o caminho para para o I de Moran é Space > Univariate Moran´s I > seleciona a variável



Vamos demonstrar apenas o I de Moran univariado. Existem outros testes de autocorrelação, quando se utiliza mais de uma variável. Para aprofundar na teoria e funções dos demais tipos de testes de autocorrelação espacial, pesquisar em Anselin (2003). O gráfico acima mostra que a votação de Bolsonaro no Paraná apresentou um coeficiente I de Moran de 0,509, ou seja, 50,9% da variação de votos em um município está associada à variação média de seus vizinhos. Este é considerado um coeficiente moderado. Além do coeficiente, o gráfico divide-se em quadrantes. Os dois mais importantes são o quadrante superior direito, que indica os municípios com alta votação e com vizinhos também com alta votação em Bolsonaro. O quadrante inferior esquerdo reúne municípios e seus vizinhos com baixa votação a Bolsonaro. Os outros dois quadrantes indicam os municípios que não apresentaram autocorrelação com seus vizinhos, seja porque a unidade apresentou baixa votação e seus vizinhos alta ou o contrário.

Para identificar quais são os municípios com autocorrelação positiva ou negativa, ainda no Geoda você pode selecionar os casos em um dos quadrantes. Isso fará com que as unidades espaciais sejam destacadas no gráfico de dispersão em vermelho e no mapa em amarelo, como indicado na figura abaixo para os municípios que compõem o quadrante superior esquerdo do gráfico de dispersão do I de Moran para votação em Bolsonaro.



A seleção de casos no gráfico de dispersão permite identificar que os municípios com autocorrelação positiva de votos em Bolsonaro encontram-se principalmente na região leste e litoral do estado, no extremo oeste e no norte do Paraná. O mapa também mostra o número de unidades selecionadas. No caso, são 171 municípios dos 399 que se encontram no quadrante superior direito do gráfico de dispersão. Se o I de Moran é um coeficiente de autocorrelação global, um complemento é aquele indicador que mostra as correlações locais. Para isso o coeficiente mais utilizado é o coeficiente Lisa.

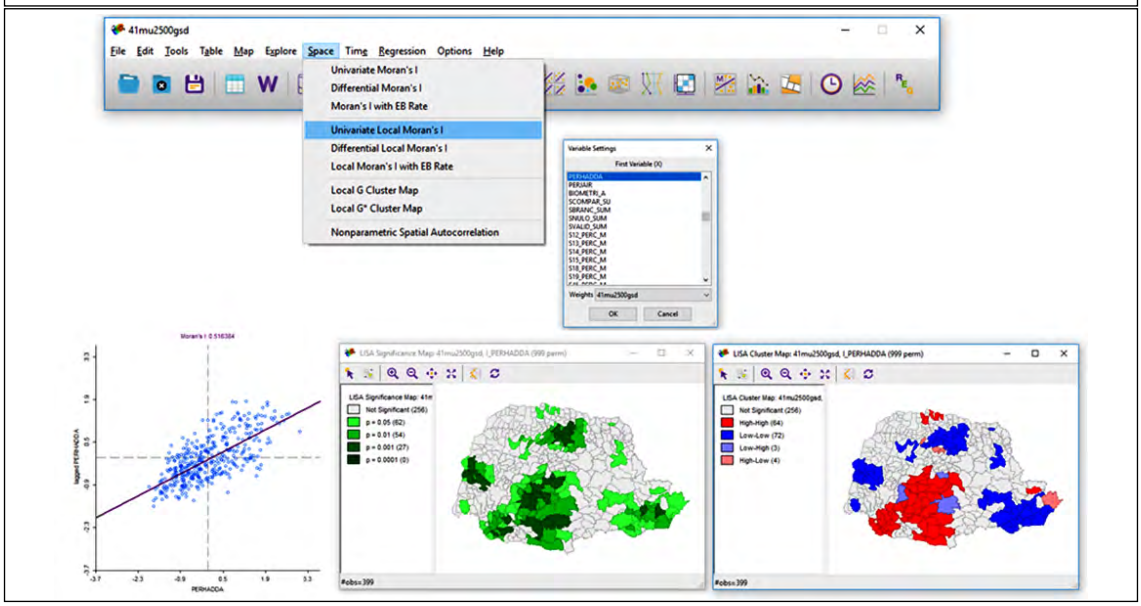
9.4.2 Coeficiente LISA para *clusters* geográficos

O coeficiente indicado para testar a similaridade de desempenho da variável analisada por vizinhança geográfica é chamado de LISA (sigla em inglês para Indicador de Associação Espacial Local), proposto por Anselin (1995). Este coeficiente é calculado a partir da decomposição de um indicador global de associação espacial, como o I Moran e, com isso, mostra a influência de unidades locais sobre a magnitude global das associações. Na prática, o que ele faz é identificar *outliers* de unidades de observação em relação à média geral da associação espacial. De acordo com Anselin (1995), o LISA dá para cada observação um indicador da extensão de *clusters* espacialmente significativos em função da similaridade de valores ao redor de cada unidade observada. Além disso, a soma do coeficiente LISA, para todas as unidades de observação, é proporcional ao indicador de associação espacial glogal (Anselin, 1995). Com isso, o LISA permite identificar a formação de *clusters* espaciais, ou seja, de conjuntos de unidades contínuas que têm coeficientes estatisticamente significativos a partir de um teste de hipóteses para associação espacial. São duas informações comple-

mentares. A primeira é se as vizinhanças apresentam formação consistente de *cluster* ou não. A segunda é se o *cluster* é positivo ou negativo. Assim, LISA apresenta cinco resultados possíveis: não significância entre vizinhos, significância para vizinhos alto-baixo, significância para vizinhos baixo-alto (estes dois são considerados estatisticamente significativos, porém, inconsistentes), significância entre vizinhos alto-alto e significância baixo-baixo (os dois últimos indicam os *clusters* estatisticamente significativos e consistentes, pois valores altos ou baixos de uma unidade se associam na mesma direção com os de seus vizinhos.

No Geoda, o teste LISA produz três *output*s. Por se tratar de um complemento ao I de Moran, ele produz um gráfico de dispersão entre a unidade de análise e a média de seus vizinhos. Além disso, são produzidos dois mapas. O primeiro deles, em tons de verde, mostra as unidades espaciais com significância estatística para formação de *clusters*. São quatro tons, um para cada nível de significância, entre p=0,05, p=0,01, p=0,001 e p=0,0001, a maior significância estatística espacial. Além deles, as unidades em branco não apresentam significância. O segundo mapa é o mais importante para o teste LISA, pois apresenta as unidades espaciais que formam *clusters* a partir de quatro cores. Vermelho escuro é o *cluster* de valores alto-alto, vermelho claro é para unidades com alto -baixo, azul escuro para *cluster* baixo-baixo e azul claro para unidades baixo-alto. Para exemplificar a formação de *clusters* pelo método LISA, usaremos a variável percentual de votos em Haddad por município do Paraná em 2018, conforme segue:

Caminho para teste LISA no Geoda: Space > Univariate Local Moran´s I > Selecione a variável (por padrão ele produzirá os três outputs descritos acima. Se quiser, pode desmarcar as caixas dos mapas ou gráfico que não precisará usar).



Os resultados apresentados acima mostram um coeficiente I de Moran de 0,516, muito próximo do apresentado pela votação do candidato Bolsonaro no estado do Paraná. O primeiro mapa mostra que dos 399 municípios, 256 não apresentaram significância estatística para LISA. Outros 63 ficaram abaixo de p=0,05, outros 54 municípios com p=0,01 e 27 municípios com p=0,001. Não houve significância estatística abaixo de p=0,0001. Percebe-se pelo mapa de significância que os *clusters* se formaram principalmente em quatro regiões. No extremo leste do estado, no sudoeste, no extremo oeste e em uma região do norte do Paraná. O que resta saber é quais *clusters* são consistentes e em que direção. O segundo mapa oferece essa informação. O *cluster* na cor vermelha escura, no sudoeste do Paraná, indica o *cluster* com unidades de análise com percentuais de votos em Haddad alto-alto (são 64 municípios nessa condição). Nos outros três *clusters* predominam municípios com percentuais de voto baixo-baixo em Haddad (são 72 municípios nessa condição). Outros três municípios apresentam *cluster* baixo-alto, na cor azul clara e mais quatro municípios, em vermelho claro, indicam desempenho alto-baixo de votos em Haddad.

Com isso, concluímos os exemplos de coeficientes de autocorrelação e de *clusters* locais para análise exploratória de dados espaciais. No próximo tópico, são apresentados os princípios gerais dos testes de regressão no Geoda, tanto da regressão linear clássica, quanto dos tipos de regressão espacial apresentados no *software*.

9.4.3 Testes de Regressão Linear para unidades espaciais no Geoda

Assim como em outros conjuntos de técnicas de análise exploratória e prescritiva, as espaciais também utilizam os princípios da regressão linear para estabelecer modelos estatísticos ajustados para descrição e possível predição de valores em função do que já está registrado. No caso da análise geográfica é acrescentado um novo fator explicativo que é a vizinhança. Ou seja, as regressões espaciais consideram, além dos efeitos das variáveis explicativas sobre a dependente, também o efeito de vizinhança controlado pelas demais variáveis para explicar as variações entre unidades de análise. Existem duas formas principais de consideração de vizinhança nas regressões espaciais – a defasagem (SAR) e o erro espacial (SEM).

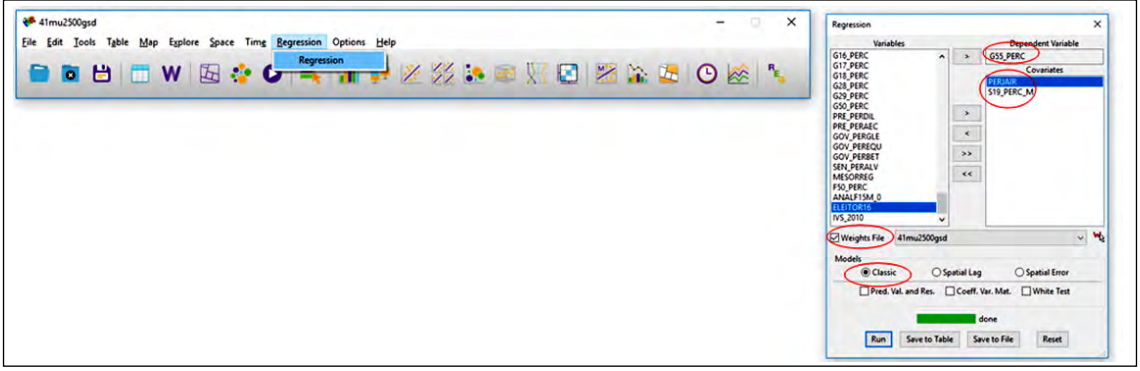
O primeiro modelo é o de defasagem espacial ou *Spatial Autoregressive Model* (SAR). Esse modelo incorpora um coeficiente autorregressivo espacial, a fim de obter o efeito de "vizinhança" do fenômeno em estudo, que captaria a forma como um fenômeno seria propagado pelas regiões próximas. O parâmetro ρ de defasagem espacial, como lembra Almeida e Guimarães (2012), tem seu valor situado no intervalo aberto entre -1 e 1. Se coeficiente for maior que a proporção na unidade, a autocorrelação será positiva. Desta forma, por exemplo, altas proporções de votos em determinado candidato ou partido nas regiões vizinhas aumenta os votos na região i. O resultado inverso, onde ρ é menor que a unidade, indica que alta proporção de votos nas regiões vizinhas diminui o valor da taxa de homicídio na região i. Se o valor de ρ for não estatisticamente significativo, não há evidências que exista autocorrelação entre as unidades de análise vizinhas. No modelo de erro autorregressivo espacial (SEM) a dependência espacial é residual, devido à estrutura autorregressiva de primeira ordem no termo de erro. Logo, utiliza-se esse modelo quando os efeitos não modelados não podem estar correlacionados com nenhuma variável explicativa da regressão.

O processo de estimação segue o procedimento recomendado por Florax e Graaf (2004), em que se deve estimar o modelo de regressão linear clássico e testar se há autocorrelação espacial. Caso exista, o procedimento seguinte será escolher entre os modelos SAR e SEM. Para isso, utiliza-se o Multiplicador de Lagrange, indicado no *output* da regressão linear. Caso o teste para ambos os modelos seja significativo, deverá ser utilizado o Multiplicador de Lagrange Robusto para verificar o maior valor estimado. Na prática, roda-se a regressão linear inicialmente para identificar os efeitos das variáveis explicativas sobre a dependente sem considerar o impacto do espaço. Em seguida, a partir dos coeficientes de Lagrange, define-se pela utilização do modelo SAR ou SEM e roda-se uma nova regressão, dessa vez, levando em conta o efeito do espaço. O resultado da segunda regressão incluirá o novo coeficiente explicativo espacial e os efeitos individuais das variáveis explicativas serão alterados em função do controle pela vizinhança.

Assim como em qualquer outro teste de regressão, são gerados dois conjuntos de coeficientes. O primeiro diz respeito ao ajustamento do modelo como um todo, tais como a estatística F, r^2 , indicadores de colinearidade e outros. O segundo é o conjunto de estatísticas dos efeitos individuais, cujos principais são o coeficiente angular (β), a estatística

"t" para cada variável explicativa e o nível de significância estatística (*p value*). Vamos usar como exemplo para o primeiro modelo de regressão, por MQO, duas variáveis de desempenho eleitoral para explicar a distribuição de votos de Ratinho Jr. nos municípios paranaenses. A primeira delas é a distribuição de votos de Bolsonaro nos municípios do estado. A segunda é a distribuição de votos do candidato a senador na chapa de Ratinho Jr., Oriovisto Guimarães (Podemos). Os objetivos são: a) identificar qual o ajustamento do modelo, ou seja, quanto que essas três variáveis juntas explicam da variação de votos de Ratinho Jr.; b) comparar os coeficientes angulares (β) de cada uma das variáveis para saber qual delas tem maior poder explicativo; e c) identificar o melhor modelos de regressão espacial a partir dos coeficientes de Lagrange. O caminho para pedir a regressão e os resultados estão no quadro a seguir. Perceba que mesmo pedindo o modelo clássico, sem considerar o efeito do espaço, é preciso inserir o arquivo com extensão ".gal" na caixa de pesos do modelo. Sem ele não será possível estimar os coeficientes Lagrange. Se o arquivo de pesos não for incluído será rodado o modelo MQO, porém, sem a apresentação de Lagrange ao final.

Caminho para regressão no Geoda: Regression > preenchem-se as caixas com a variável dependente e as independentes > carrega o arquivo de peso espacial (W) > escolhe o modelo a ser usado (clássico MQO, defasagem ou erro espacial).




```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  41mu2500gsd
Dependent Variable  :    G55_PERC  Number of Observations:  399
Mean dependent var  :     56.4249  Number of Variables   :    3
S.D. dependent var  :       10.68  Degrees of Freedom    :  396

R-squared           :    0.214563  F-statistic           :     54.0891
Adjusted R-squared  :    0.210596  Prob(F-statistic)     :1.70616e-021
Sum squared residual:     35745.9  Log likelihood        :    -1462.95
Sigma-square        :     90.2675  Akaike info criterion :     2931.91
S.E. of regression  :     9.50092  Schwarz criterion     :     2943.88
Sigma-square ML     :     89.5888
S.E of regression ML:     9.46513
```

```
------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------
          CONSTANT       38.0521        2.16456        17.5796     0.00000
           PERJAIR      0.115371       0.050911        2.26613     0.02398
        S19_PERC_M      0.554197      0.0776341        7.13857     0.00000
------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   11.460171
TEST ON NORMALITY OF ERRORS
TEST                   DF           VALUE             PROB
Jarque-Bera             2             5.5843          0.06129

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                   DF           VALUE             PROB
Breusch-Pagan test      2            58.9217          0.00000
Koenker-Bassett test    2            45.7098          0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : 41mu2500gsd
   (row-standardized weights)
TEST                          MI/DF        VALUE          PROB
Moran's I (error)             0.2811        9.3865        0.00000
Lagrange Multiplier (lag)        1          54.5863        0.00000
Robust LM (lag)                  1           6.3803        0.01154
Lagrange Multiplier (error)      1          82.9721        0.00000
Robust LM (error)                1          34.7661        0.00000
Lagrange Multiplier (SARMA)      2          89.3524        0.00000
================================ END OF REPORT ================================
```

A primeira parte do *output* de resultados mostra os coeficientes de ajustamento geral do modelo. A média da variável dependente (56,42) é o percentual médio de votos obtidos por Ratinho Jr. nos municípios do estado. O $r^2 = 0,214$ indica que o ajustamento do modelo é de 21,4%, ou seja, as variáveis independentes explicam 21,4% das variações da votação de Ratinho Jr. nos municípios. Não é um ajustamento alto. Mas a estatística F = 54,08 (*p-value* = 1.70616e-021) mostra significância estatística em pelo menos uma das variáveis explicativas. O critério de Akaike é usado comparativamente entre modelos. Aquele que tiver o menor valor de Akaike será o modelo com melhor ajuste, portanto, sem um segundo modelo para comparação, ele não tem utilidade. De qualquer maneira, considerando o baixo (r^2), é possível esperar um valor alto de Akaike, pois o modelo está explicando pouca variação da variável dependente.

Na segunda parte dos resultados, aparecem os coeficientes individuais das variáveis explicativas. Percebe-se que os dois apresentam significância estatística (Probabilidade < 0,050). Comparando os coeficientes angulares de cada uma delas, percebemos que a mudança de 1% de votos para Bolsonaro implica em 0,11% a mais de voto de Ratinho Jr (β

= 0,115371), enquanto a mudança de 1% a mais de votos em Oriovisto (S19_perc_M) gera 0,55% a mais de voto em Ratinho Jr. (β = 0,554197). Ou seja, ainda que as duas tenham efeitos significativos do ponto de vista estatístico e sejam positivas, a variação de votos em Oriovisto explica cerca de cinco vezes mais as variações de Ratinho Jr. que as de Bolsonaro.

O terceiro conjunto de informações diz respeito à qualidade das variáveis explicativas para o modelo. O diagnóstico de colinearidade indica se as variáveis independentes são colineares em suas variações. A colinearidade é um problema para modelos preditivos, pois ela aumenta artificialmente os efeitos explicativos. Para que não seja colinear, o coeficiente do diagnóstico precisa estar abaixo de 10,0. No nosso exemplo ele é de 11,46, mostrando existência de colinearidade, o que é esperado, pois as votações de Oriovisto e de Bolsonaro tenderam a ser "casadas" no Paraná. De qualquer maneira, com esse diagnóstico seria perigoso fazer predições de votação a partir das variáveis explicativas. O segundo coeficiente é o teste de normalidade de Jarque-Bera para erros. Espera-se em um modelo de regressão que as distribuições dos erros do modelo se aproximem do formato de uma curva normal. Como o resultado indica um *p-value* > 0,050 (probabilidade = 0,061), podemos rejeitar a hipótese de que os erros não seguem uma curva normal e aceitar que eles se aproximam da normalidade. Por fim, os testes de heteroquedasticidade indicam que a distribuição dos erros apresentam formato homoquedástico – o que é um pressuposto importante para os modelos de regressão. Como no modelo os dois testes apresentam *p-value* > 0,050, podemos rejeitar a hipótese de heteroquedasticidade e assumir que o formato é homoquedástico e, portanto, não quebram o pressuposto da regressão.

O último conjunto de informações diz respeito ao diagnóstico de dependência espacial. Só foi gerado porque incluímos no modelo o arquivo de peso espacial (W). A partir dele podemos tomar duas decisões. A primeira é se existe dependência espacial. Para isso, olhamos para os indicadores de probabilidade. Se eles estiverem com *p-value* < 0,050 existe dependência espacial. Como é o nosso caso para a maior parte os indicadores. A segunda é qual o melhor modelo de dependência espacial deve ser usado. Se o de defasagem (lag) ou o de erro. Basta olhar para o maior valor de coeficiente dos indicadores de Lagrange. No nosso caso, Lagrange lag = 54,58 e Lagrange erro = 82,97. Assim, devemos usar o modelo de erro espacial. Se os valores estivessem muito próximos, passa-se à análise das diferenças de Lagrange robusto, o que não foi necessário nesse caso.

Tomada a decisão de recalcular o modelo de regressão com dependência espacial pelo erro, o quadro a seguir mostra os resultados do novo modelo.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : 41mu2500gsd
Spatial Weight      : 41mu2500gsd
Dependent Variable  :    G55_PERC  Number of Observations:  399
Mean dependent var  :   56.424862  Number of Variables   :    3
S.D. dependent var  :   10.679998  Degrees of Freedom    :  396
Lag coeff. (Lambda) :    0.517357

R-squared           :    0.370605  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-1430.221324
Sigma-square        :     71.7903  Akaike info criterion :     2866.44
S.E of regression   :     8.47292  Schwarz criterion     :     2878.41

------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------
          CONSTANT       37.2917        2.51541        14.8253     0.00000
           PERJAIR     0.0812037      0.0569074        1.42694     0.15360
         S19_PERC_M      0.686573      0.0822934        8.34298     0.00000
            LAMBDA      0.517357      0.0578621        8.94121     0.00000
------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      2        63.5007     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : 41mu2500gsd
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1        65.4671     0.00000
================================ END OF REPORT ================================
```

Os resultados gerais do modelo indicam uma melhora do ajustamento, ainda que no geral ele continue baixo, com $r^2 = 0,370$, ou seja, antes de considerar a dependência espacial o percentual de variação explicada era de 21%, com a dependência espacial a explicação sobe para 37% da variação. O critério Akaike também diminuiu em relação ao modelo anterior, confirmando a melhora do ajustamento. O teste de diagnóstico de heteroquedasticidade indica que podemos continuar rejeitando a hipótese da quebra de pressuposto da homoquedasticidade. Assim como o teste de razão de verossemelhança (prob = 0,000). Ou seja, em geral a consideração da dependência espacial pelo erro melhorou o modelo explicativo.

Mas, as principais mudanças encontram-se nos efeitos individuais das variáveis. A primeira delas é que ao considerarmos a dependência espacial, a votação de Bolsonaro deixa de apresentar significância estatística para as variações de voto em Ratinho Jr (pro-

babilidade = 0,15360). Isso quer dizer que o efeito de vizinhança da votação de Ratinho Jr. é maior que o efeito dos votos de Bolsonaro. O coeficiente de dependência espacial nos modelos por erro é representado por Lambda. No caso, a vizinhança apresenta um efeito de 0,5173 sobre a votação de Ratinho Jr. No entanto, o maior efeito individual continua sendo a votação do candidato ao senado, Oriovisto Guimarães (S19_perc_M), com significância estatística e coeficiente β = 0,6865. A conclusão é que o modelo de ajustamento por dependência espacial melhorou a capacidade explicativa da variação de votos em Ratinho Jr. principalmente porque ele demonstrou que o efeito de vizinhança foi maior que a votação de Bolsonaro, porém, menor que a de Oriovisto, para Ratinho Jr. Em outras palavras, nos municípios do Paraná os votos de Ratinho Jr. e Oriovisto tenderam a variar juntos, inclusive quando se considera a dependência espacial.

Neste capítulo procuramos apresentar ao leitor uma introdução às técnicas de análise geográfica, seus princípios, pressupostos e uma aplicação ao *software* Geoda. Por óbvio, muitas coisas não foram abordadas, seja em relação aos princípios e teoria da análise espacial, seja em relação às funcionalidades do Geoda. Minha sugestão é que o leitor aprofunde os conhecimentos a partir de seus interesses específicos, usando os manuais e tutoriais indicados ao longo do capítulo ou além dele. No Geoda, por exemplo, existe uma área (Explore) para geração de estatísticas e gráficos descritivos básicos como histograma, boxplot, gráfico de dispersão e outros que não foram apresentados aqui. Também é possível alterar o banco de dados, incluindo, calculando ou excluindo variáveis ou conjuntos de casos a partir do “table”. Por fim, o banco de dados usado no capítulo e nos exercícios a seguir apresenta um conjunto grande de variáveis para serem exploradas pelos que pretendem continuar usando a geografia como variável explicativa em suas pesquisas.

9.5 Referências bibliográficas do Capítulo IX

Almeida, E., & Guimarães, P. (2012). *Income convergence and infrastructure in Brazil: a spatial multilevel approach*. Mimeo, Department of Economics, UFJF.

Alkmim, Antonio. (2014). De Brizola a Cabral. *De Collor a Dilma: a geografia do voto no Rio de Janeiro de 1982 a 2010*. Rio de Janeiro: Ed. Puc-Rio.

Anselin, Luc., & Rey, Serge. (1991). Properties of tests of spatial dependence in linear regression models. Geographical *Analysis*. 23(2), 112-131.

Anselin, Luc. (1995). Local Indicators of Spatial Association – Lisa. *Geographical Analysis*. 27(2), 93-115.

Anselin, Luc. (2003). Geoda User´s Guide. *Center of Spatially Integrated Social Science*. Ilinois: University of Ilinois.

Books, John W., & Prysby, Charles L. (1991). Political behavior and the local context. New York: Prager.

Braga, Maria do S., & Rodrigues-Silveira, Rodrigo. (2011). Competição partidária e territoriedade do voto: mudanças na distribuição do padrão espacial do voto para presidente da república do Brasil. *Paper apresentado no XXXV Encontro anual da Anpocs*, Caxambu-MG.

Florax, R. J. G. M., & Graaff, T. (2004). The performance of diagnostic tests for spatial dependence in linear regression models: a meta-*analysis* of simulation studies. In Anselin, L., Florax, R. J. G. M. e Rey, S. J. (ed.). *Advances in spatial econometrics*. Springer, New York.

Johnston, R. J. (1983). Spatial continuity and individual variability: a revied of recent work on the geography of elecotral change. *Electoral Studies*, 2(1), 53-68.

Johnston, R. J. *et al.* (2001). Social location, spacial locations and voting at the 1997 British general election: evaluating the sources of Conservative support. *Political Geography*, 20, 85-111.

Miller, W. L. (1977). *Electoral dynamics*. London: Macmillan.

Potter, Joshua D., & Olivella, Santiago. (2015). Electoral strategy in geographic space: accounting for spatial proximity in district-level party competition. *Electoral Studies*, 40, 76-86.

Rodrigues-Silveira, Rodrigo. (2013). Representación espacial y mapas. *Cuadernos Metodológicos*, 50. Madrid: Centro de Investigaciones Sociológicas.

Soares, Glaucio A. D. (1973). Desigualdades eleitorais no Brasil. *Revista de Ciência Política*: Rio de Janeiro, 7(1), 25-48.

Terron, S. L., & Soares, G. A. D. (2010). As bases eleitorais de Lula e do PT: do distanciamento ao divórcio. *Opinião Pública*, 16(2), 310–337.

Terron, Sonia. (2012). Geografia eleitoral em foco. *Revista Em Debate*, 4(2), 2, 8-18.

Zavala, Rita G. B. (2012). Génesis dela geografia electoral. *Revista Espacialidades,* 2(1), 81-95.

9.6 Exercícios propostos do Capítulo IX

Faça o download da pasta zipada de arquivos “BDCAP9_AE” em (http://www.filedropper.com/bdcap9ae). A pasta zipada está no site File Dropper. Para ter acesso a ela, basta copiar o endereço acima, colar na barra de endereço do *browser* e rolar a página até a área onde consta “Download”. Com um clique o arquivo será transferido para seu computador. Salve em um local no disco e descompacte a pasta. Depois disso, abra o programa Geoda que já deve estar instalado no seu computador. Na área “input” da caixa de diálogo, introduza o arquivo “shp” da pasta que acabou de descompactar. Abra o shape no Geoda e produza os seguintes mapas e testes:

9.6.1 Gere o mapa por quebras naturais para a votação de Ciro Gomes (PDT) no Paraná (variável “PERCIRO”) e gere o mapa de Desvio Padrão para a mesma variável. Interprete as diferenças entre os dois mapas.

9.6.2 Gere o gráfico de dispersão de I de Moran para a variável “PERCIRO” e interprete o coeficiente.

9.6.3 Gere o gráfico e mapas do coeficiente LISA para a variável “PERCIRO” em municípios do Paraná e interprete os mapas de significância e de *clusters* de LISA.

9.6.4 Identifique qual o melhor modelo de dependência espacial para uma regressão onde a variável dependente é “PERCIRO” e as variáveis explicativas são votação para deputado federal do PDT por município do Paraná (F12_PERC) e votação em Ratinho Jr. (G55_PERC). Interprete os resultados.

Capítulo X

Análise de Séries Temporais

Você já arremessou dardos tentando acertar um ponto central? Quantas vezes você acertou? Em quantas o dardo atingiu pontos distribuídos aleatoriamente ao redor do alvo? (Gujarati, 2006).

Analisar a distribuição de casos em pontos distintos do tempo exige um conjunto próprio de técnicas estatísticas. A análise de séries temporais, por origem, quebra o pressuposto da independência entre as observações, pois elas estão arranjadas no tempo. Sendo o tempo a variável explicativa, uma ocorrência em determinado tempo é o resultado do que aconteceu no tempo anterior mais um efeito externo qualquer, que pode ser aleatório ou não. Analisar fenômenos e variações ao longo do tempo usando técnicas clássicas de regressão tende a resultar em relações espúrias ou regressões cujos resultados não fazem sentido. Outra característica das séries temporais é que o conjunto de técnicas está em permanente evolução.

O objetivo deste capítulo é apresentar os conceitos básicos das técnicas de análise de séries temporais, com prioridade para a aplicação a análises descritivas de séries passadas – que possam explicar variações dos fenômenos políticos no tempo – e não para a previsão de valores futuros – que tende a ser o objetivo principal desse tipo de técnica, quando aplicada às variáveis econômicas. Como nos demais capítulos do livro, optamos por apresentar as técnicas de séries temporais a partir de um *software* livre, com

código aberto, específico para esse tipo de análise. Trata-se do Gretl, que é acrônimo de (*Gnu Regression, Econometrics and Time-series Library*)[1]. No anexo a este capítulo, há um passo a passo para instalação do programa e carregamento de banco de dados.

10.1 Fundamentos

As análises de séries temporais permitem modelar de forma dinâmica alterações ocorridas em função do tempo para explicar variações do passado, com explicações sobre resultados, ou predizer possíveis valores no futuro, a partir da inferência. Ao contrário de outras técnicas de análise que excluem de seus modelos o caráter dinâmico dos dados por trata-los como violação de pressupostos, a análise de séries temporais incorpora a dinâmica temporal no modelo analítico, que pode ser: i) modelo estacionário univariado; ii) modelo estacionário multivariado, ou iii) modelo não estacionário uni ou multivariado (Pevehouse & Prozek, 2008). Como veremos mais adiante, os modelos (i) e (ii) são necessários para fazer predição temporal e por isso uma parte considerável do conjunto de técnicas de séries temporais se concentra em como transformar séries de dados em valores estacionários.

Análise de séries temporais, seja passado ou presente, requer um conjunto de técnicas específicas. Os valores estão sincronizados com as unidades de tempo. Isso exige homogeneidade nas observações ao longo do tempo. Nos modelos de regressão clássica pelo método dos Mínimos Quadrados Ordinários (MQO), assume-se que não existe autocorrelação entre os termos, ou seja, os valores de uma variável não são relacionados entre si nos casos analisados. Isso não ocorre quando a variável explicativa é o tempo, pois por princípio um evento em determinado momento do tempo carrega uma "memória" do que aconteceu no passado e terá influência parcial naquilo que virá no futuro. Praticamente todos os processos sociais, quando analisados no tempo, carregam uma inércia do passado e são dirigidos por forças históricas. Assim, conceitualmente,

[1] Gnu (*general public license*) é a sigla de um projeto iniciado em 1982 por Richard Stallman para criar um sistema operacional totalmente livre. Dentro do projeto foi criada a FSF, sigla em inglês para Fundação do *Software* Livre. Para acessar o manual do *software* em português o link é: http://gretl.sourceforge.net/gretl_portugues.html

uma série temporal pode ser dividida em duas partes principais: um componente de tendência, mais regular no tempo, e um componente de variações sazonais ou cíclicas, que é mais irregular.

O objetivo de uma série temporal normalmente é predizer o comportamento de um fenômeno com mudança ao longo do tempo. Como lembra Gujarat (2006), para verificar se uma previsão é válida, antes é preciso saber se os dados são ou não estacionários no tempo. Só depois é que se fazem os testes de causalidade e predição de valores faltantes. As variações podem ser evolutivas ou estacionárias. Ela é evolutiva quando a média e a variância mudam ao longo do tempo. Estacionária é aquela cujas média e variância se mantêm estáveis ao longo do tempo.

São dois os principais objetivos em uma análise de série evolutiva ou estacionária. O primeiro é descrever o que aconteceu no passado, identificando se os dados respondem a determinado padrão de comportamento. O segundo é, uma vez definido esse padrão, tentar predizer o comportamento futuro da série. Assim, toda série temporal é uma função do tempo, onde a variável dependente é a própria série e o tempo é a variável explicativa. Ou seja, aqui o tempo não é apenas um suporte, mas sim uma explicação do comportamento.

Uma das metas iniciais mais comuns desse tipo de análise é a identificação de tendências. Uma tendência temporal é definida como a correlação dos desvios de uma média ao longo do tempo. Essa correlação de desvios pode ser usada para testar similaridades maiores ou menores em determinados períodos. Se existir alguma periodicidade aparente na série, é provável que a correlação das diferenças apareça mais em mudanças ocorridas em curtos espaços de tempo (Pevehouse & Prozek, 2008). Aqui, surge uma primeira diferença importante: Efeitos de Curto e Efeitos de Longo Prazo. Técnicas de análise de séries temporais permitem identificar, separar e oferecer tratamentos distintos para cada um dos efeitos. Outra diferença importante é entre uma série que apresenta valores constantes ao longo do tempo e aquelas cujas médias tendem a se alterar, seja crescendo ou diminuindo. No primeiro caso, temos o que se chama de Médias Estacionárias, o que é importante para predição temporal.

Uma série temporal possui quatro componentes:

- Tendência (T): é o componente que altera o comportamento da série a longo

prazo e em uma mesma direção. Necessita de muitas observações para ser identificada. Uma tendência pode ser linear, exponencial, parabólica, logística, etc. A presença de tendência mostra que a série é evolutiva e não estacionária.

- Variações sazonais (VS): movimentos da série que se repetem periodicamente. Normalmente causada por ordenação do tempo, como ocorrências em meses ou dias específicos. São movimentos de curto prazo.

- Variações cíclicas (VC): normalmente acontece em variáveis econômicas, que apresentam ciclos expansivos e recessivos. São movimentos de médio prazo.

- Variações residuais ou acidentais (R): é um comportamento que não responde a nenhum dos fenômenos anteriores, mas de fatores aleatórios que influenciam de forma isolada e temporária a série.

A interação desses quatro fatores gera a série temporal, que é representada pela fórmula aditiva:

$$y_t = T_t + VE_t + C_t + R_t$$

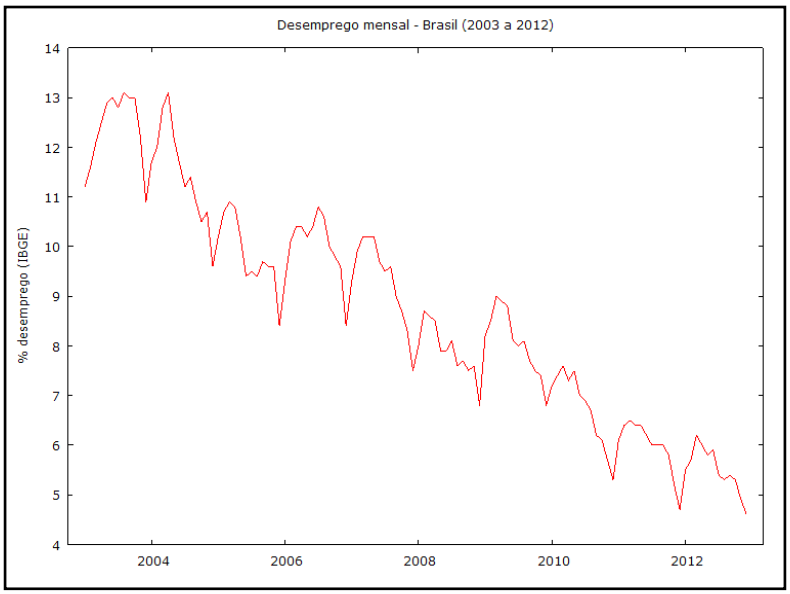
Além do efeito direto do tempo, a análise de séries temporais permite o enfoque causal, onde as variações de uma série podem ser explicadas por outras séries temporais. O primeiro fator a ser analisado em séries temporais é a tendência. Uma tendência pode ser de dois tipos: Determinista ou Estocástica. Uma tendência determinista ocorre quando seus movimentos não são aleatórios, ou seja, apresentam uma determinação de sentido marcada ao longo do tempo. Já a tendência estocástica não apresenta direção claramente definida ao longo do tempo. Para identificar se uma tendência é determinista, pode-se regredir os valores da variável a alguma função linear temporal e analisar os resíduos para verificar se eles apresentam alguma direção que comprometa a regularidade de distribuição ao longo do tempo (Pevehouse & Prozek, 2008). Veremos mais adiante como controlar os efeitos determinísticos de tendências temporais. Por agora, vamos começar a identificar os componentes desse tipo de série.

O primeiro passo da análise de séries temporais é a produção de gráficos.

O gráfico permite identificar inicialmente a existência de tendência na série. O importante na identificação das séries é que o período de tempo seja amplo o suficiente para

evitar que se confundam variações cíclicas com tendências. As tendências podem ser crescentes ou decrescentes, lineares ou não lineares. Por exemplo, o gráfico a seguir mostra a série temporal da taxa de desemprego no Brasil entre janeiro de 2003 e dezembro de 2012.

Gráfico 10.1. Série temporal Taxa Desemprego Brasil



A análise visual da série indica a existência de dois elementos. Há uma clara tendência de queda ao longo do tempo. Além disso, percebem-se movimentos cíclicos sazonais ao longo de toda a série. A seguir, são apresentados os rudimentos conceituais das principais técnicas de análise temporal, na ordem proposta pelos autores especializados nesse conjunto de técnicas.

10.2 Médias Móveis

A mensuração das médias móveis serve para “suavizar” a série a partir das médias de valores de períodos fixos de tempo. Com isso, eliminam-se movimentos de curso e médio prazo e irregularidades de fatores não controlados. Com as médias móveis, são retirados três dos elementos da série, restando apenas a tendência. A ideia é que com as médias móveis se elimina dispersão e variabilidade da série, motivadas por fatores conjunturais e esporádicos. O cálculo é o de médias aritméticas de um conjunto

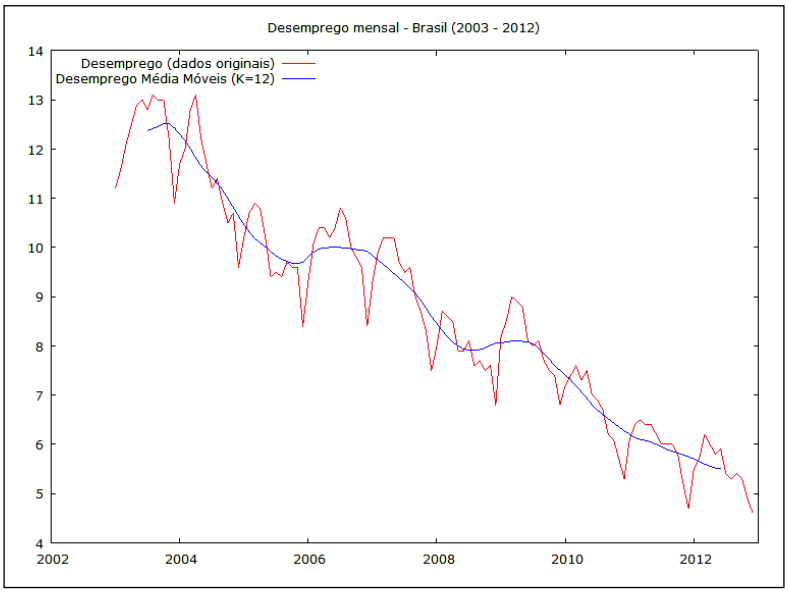
(K) de valores consecutivos. Por exemplo, imagine aplicar médias móveis com K = 5 para uma série com sete observações (Yt). Seria representado por:

$$Y_2 = \frac{Y_0 + Y_1 + Y_2 + Y_3 + Y_4}{5} \quad Y_3 = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5} \quad Y_4 = \frac{Y_2 + Y_3 + Y_4 + Y_5 + Y_6}{5}$$

Perceba que os novos valores de Y* não equivalem a nenhum valor original da série. Quanto maior for K, mais suavizada será a série. Um K muito grande leva a uma perda de informações que pode dificultar análises posteriores, pois a série pode se suavizar demais, ocultando movimentos de tendência e reduzindo demais o número de pontos da série. Por outro lado, se K for muito reduzido, não cumprirá a função de retirar as varrições cíclicas e estacionais.

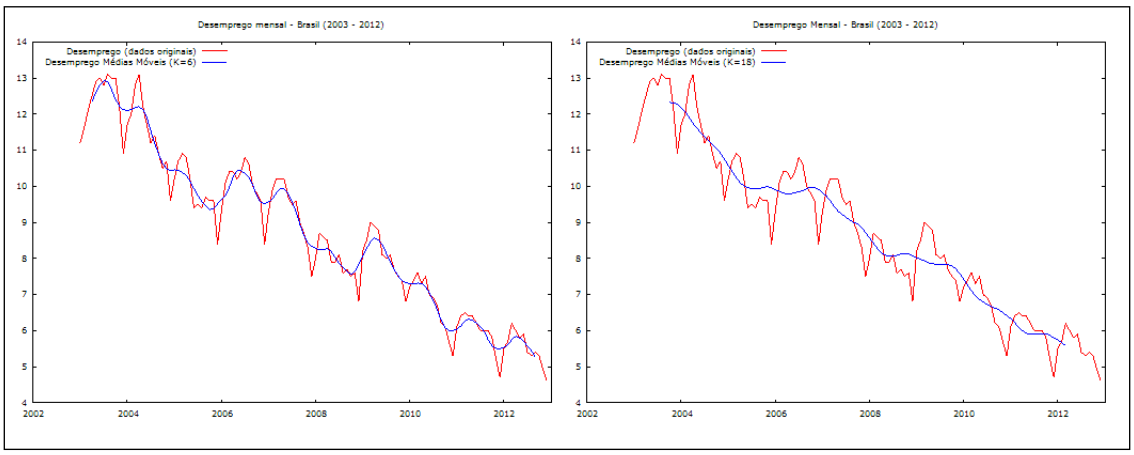
O uso da média móvel permite a identificação de tendências livres de variações estacionais ou cíclicas, mas não permite a previsão de valores futuros, pois não estabiliza a média e as variâncias ao longo do tempo. Aplicando as médias móveis para o exemplo anterior da taxa de desemprego mensal, temos a suavização das variações sazonais indicadas pela linha azul do gráfico. Como a unidade temporal é o mês e consideramos que existe sazonalidade anual para o desemprego, aplicamos as médias móveis para doze meses (K = 12).

Gráfico 10.2. Taxa de Desemprego original e Série suavizada



Para demonstrar o efeito de K sobre a suavização gerada pelos cálculos de médias móveis, os dois gráficos a seguir indicam a mesma série, porém com número de unidades espaciais distintas no cálculo. O gráfico da esquerda mostra a suavização por K = 6 e o da direita por K = 18. Perceba que, na média móvel para seis unidades de tempo, a suavização dos pontos de sazonalidade é menor, porém, apresenta um número maior de unidades de medição no tempo. Já a média móvel para 18 unidades de tempo apresenta a curva mais suavizada de dados, retirando praticamente todos os efeitos sazonais e deixando apenas a tendência decrescente no tempo. Porém, com isso, o número de unidades temporais diminui.

Gráfico 10.3. Comparação entre duas séries de médias móveis com K distintos

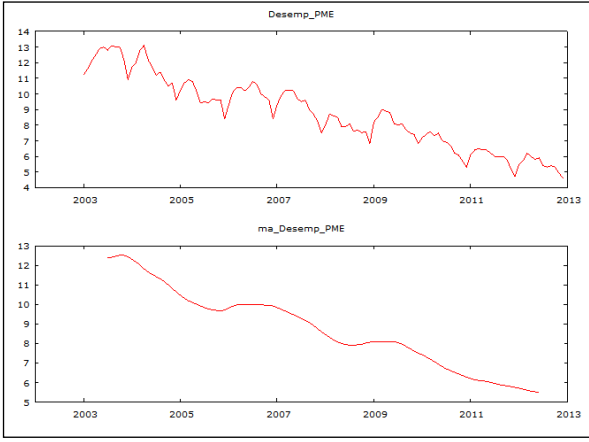


Uma informação importante para a análise dos gráficos de médias móveis acima é que as médias foram centradas. Ou seja, os dados retirados da série para produzir a primeira média não estão concentrados no início da série, mas divididos entre o início e o fim dela. O método das médias móveis é apenas a forma mais simples de suavização de uma série temporal com retirada de efeitos sazonais, cíclicos ou acidentais. Existem outras, como, por exemplo, o alisamento exponencial, que não será apresentado aqui.

Um componente sazonal, que se repete de forma sistemática em uma série, dificulta a comparação direta entre valores sucessivos da mesma, já que a média é influenciada pela estacionalidade. Para evitar esse problema, faz-se a dessazonalização da série com a chamada correção sazonal. Para isso, é preciso em primeiro lugar isolar o componente sazonal. Uma das formas mais comuns é isolar os efeitos das médias

móveis via procedimento conhecido por “razão das médias móveis”. Também existe a suavização exponencial. O exemplo a seguir é de suavização da série desemprego no Brasil entre 2003 e 2012 pelo método das médias móveis no *software* Gretl.

Gráfico 10.4. Estatísticas descritivas de série original e suavizada por k=12



Estatísticas Descritivas, observações 2003:01 - 2012:12
Variável Desemprego (120 observações válidas)

Média	Mediana	Mínimo	Máximo
8,7258	8,6500	4,6000	13,100
Desv. Padrão	C.V.	Enviesam.	Interv. IQ
2,2830	0,26164	0,14437	3,6000

Estatísticas Descritivas, observações 2003:01 - 2012:12
Média Móvel Desemprego (108 observações válidas)

Média	Mediana	Mínimo	Máximo
8,7186	8,5250	5,5042	12,517
Desv. Padrão	C.V.	Enviesam.	Interv. IQ
2,0062	0,23011	0,12320	2,8948

Os gráficos acima mostram a tendência original de desemprego e a média móvel K = 12 centralizada para desemprego (ma_desemprego). Para entender o processo de suavização, perceba as diferenças nas estatísticas descritivas das duas séries. A série de médias móveis tem um N menor, pois foram usados K = 12, o que reduz o N = 120 original para N = 108 observações válidas. As médias são muito próximas, em torno de 8,7% ao mês. A mediana da série suavizada é um pouco mais baixa que a original, porém, as principais diferenças estão nas medidas de dispersão. Os valores mínimo e máximo da suavizada são mais próximos entre si do que na variável original. Todas as demais medidas de dispersão são menores na suavizada do que na original, pois este é o objetivo da suavização da série, evitar os pontos distantes da média.

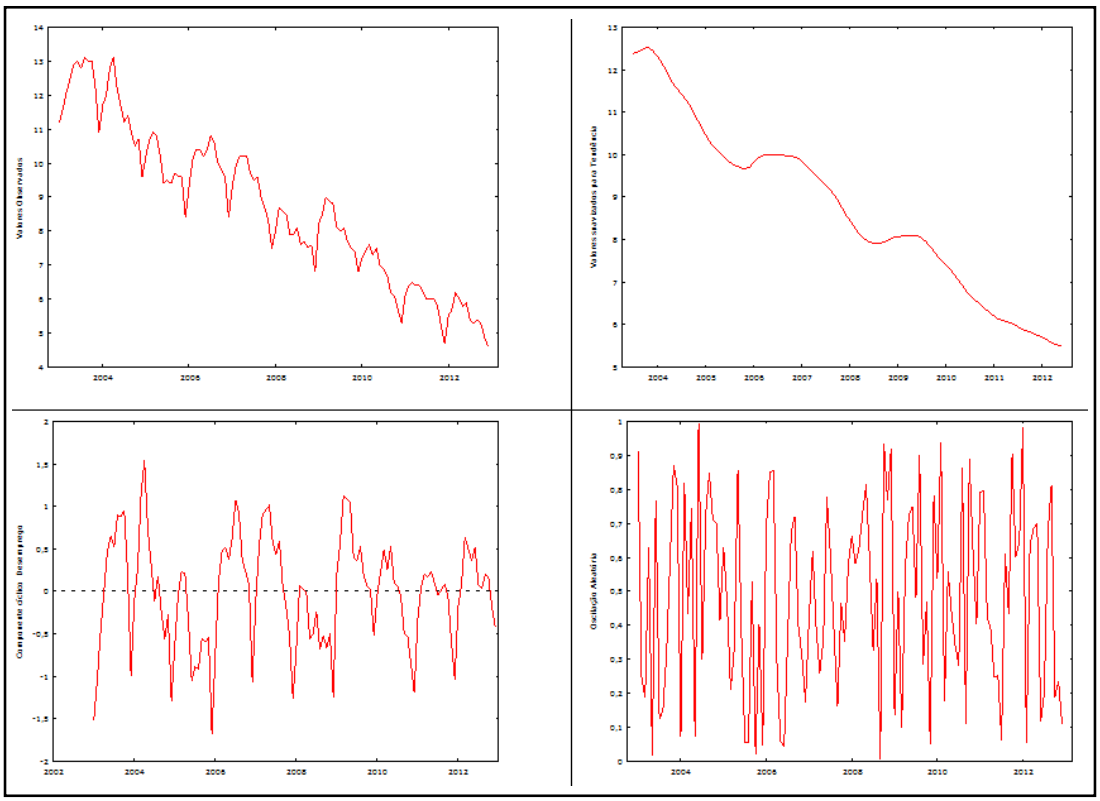
O segundo passo, uma vez estabelecido o componente de tendência, é utilizar um cociente para identificar os outros componentes de estacionalidade e variações acidentais para continuar com as análises da série. O objetivo principal é encontrar uma função de tempo que minimize a soma dos quadrados dos erros, com isso, obtém-se uma série temporal com a mesma média e variância estável. A função mais utiliza é a linear, representada pela fórmula:

$$Y_t = Y_t^* + e_t += a + bt + e_t$$

Onde: Yt é a série original, decomposta na tendência Y*t e outros componentes residuais, onde b é a variação média dos períodos Y. T é o tempo cronológico, enquanto α é o coeficiente da constante. O et é um fator de erro aleatório no tempo. Outras formas de modelar uma tendência temporal são a polinomial, exponencial, logarítmica recíproca ou logística. Vejamos a seguir os resultados de diferentes formas de decomposição de uma série:

Decomposição da série temporal:

Gráfico 10.5. Diferentes formatos de séries decompostas



Onde: Em 1926, o estatístico Undy Yule percebeu que o método utilizado até então para a diferenciação de variância ocultava oscilações que eram interessantes para a análise. Ele percebeu que o objetivo não era isolar os resíduos randômicos, mas sim as oscilações de durações diferentes para, em seguida, com método generalizado, poder dar atenção às oscilações que não eram facilmente percebidas. Ele desenvolveu um filtro para os dados que são removidos. Esse filtro basicamente é um processo autoregressivo (AR). A AR adota um par de pontos na série temporal. Com isso, a relação entre variáveis em função do tempo deixa de ser uma relação causal direta, mas sim um processo autoregressivo de termos sucessivos que são relacionados entre si (Pevehouse & Prozek, 2008).

10.3 Funções de Autocorrelação (FAC) e Autocorrelação Parcial (FACP)

As chamadas: Função de Autocorrelação (FAC) e Função de Autocorrelação Parcial (FACP) são resultados do princípio autoregressivo de Yule. Essas funções normalmente são representadas na forma de gráficos de defasagens temporais para identificar a quebra do pressuposto da estacionaridade da série. Uma vez identificada a série estacionária, é possível aplicar o método autoregressivo de médias móveis integradas (ARIMA), que foi proposto inicialmente por Box e Jenkins (1976). A proposta deles divide a análise em três estágios: (a) interpretação visual dos gráficos FAC e FACP, (b) estimação de coeficientes ajustados no modelo e (c) verificação e diagnóstico de valores preditos.

O objetivo da decomposição é transformar a série em estacionária, o que permitirá fazer predições sem romper o pressuposto da manutenção de mesma média e variância ao longo de toda a série. Para tanto, utiliza-se o conceito de função de autocorrelação.

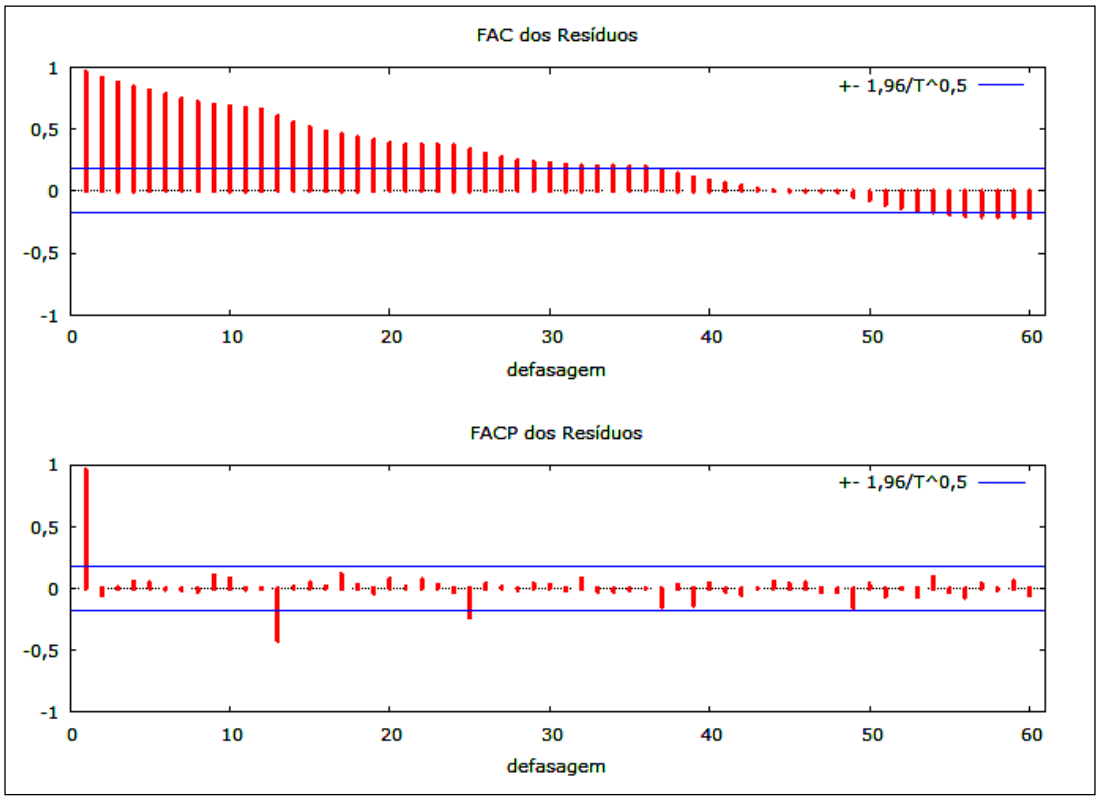
A função de autocorrelação é usada para medir a correlação entre valores de uma série distantes entre si por um período de tempo K e produz um Coeficiente de autocorrelação como resultado. Normalmente, o coeficiente é de ordem 1, chamado de r1, mas ele pode ser distanciado em ordem 2, 3 ou qualquer outra ordem K de tempo. Com o coeficiente é possível calcular o erro padrão e o intervalo de confiança para o coeficiente de autocorrelação.

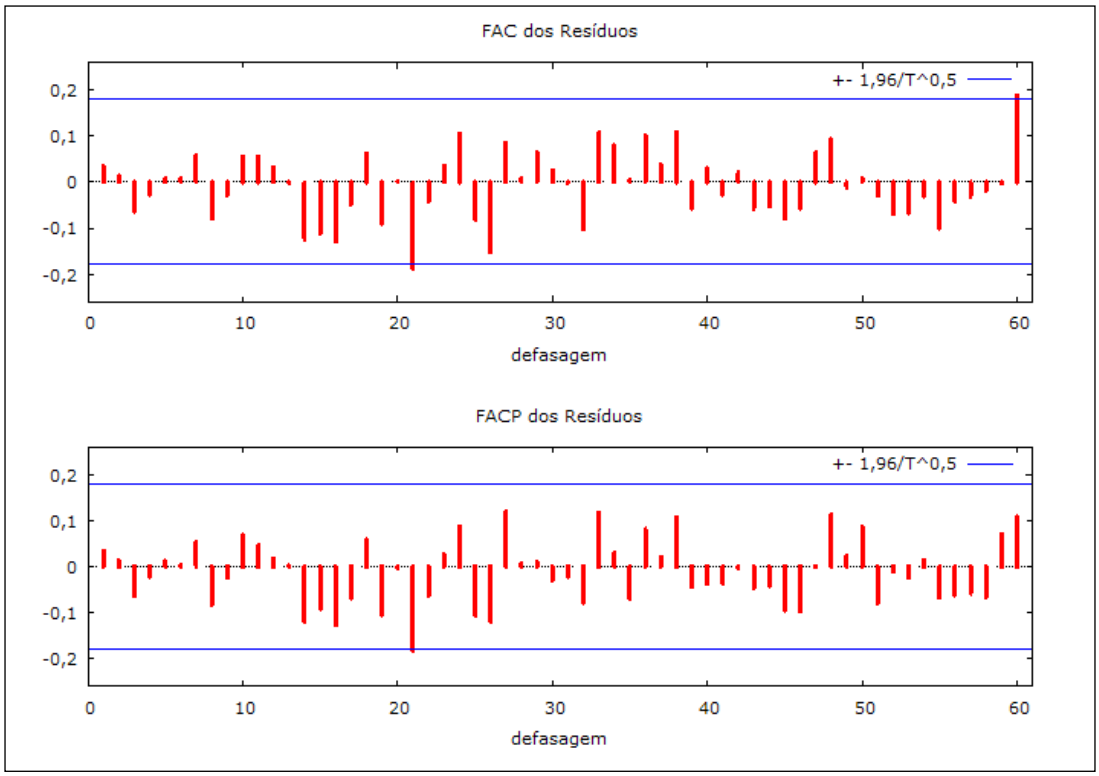
Uma função de autocorrelação nada mais é do que o conjunto de coeficientes de ordem 1 até o máximo (não pode ultrapassar a metade dos valores observados) e permite estudar a sazonalidade da série. Diz-se que um coeficiente de autocorrelação para um retorno igual ao período sazonal deve ser significativamente diferente de zero.

Uma função complementar à de autocorrelação é a de autocorrelação parcial, que é um coeficiente parcial de ordem K. Ele é calculado a partir de pares de valores separados em mesmas distâncias no tempo e eliminando o efeito que é gerado pela correlação por retornos anteriores/maiores que K. O gráfico a seguir mostra a função de autocorrelação parcial com intervalos de confiança indicados para mostrar a existência de sazonalidades. Espera-se que, após o primeiro retardo, todos os demais valores fiquem abaixo do intervalo de confiança. Nesses casos, não haverá ocorrência de sazonalidade na série transformada.

Seguindo o exemplo da série de desemprego no Brasil entre 2003 e 2012, os dois correlogramas a seguir mostram os gráficos de autocorrelação (FAC) e autocorrelação parcial (FACP). Como já vimos antes, a série natural apresenta tendência, logo, haverá autocorrelação entre os valores no tempo. É o que indicam os gráficos da direita, produzidos a partir dos valores reais. Os gráficos têm 60 retardos por ser esse o valor da metade do total de observações (120 meses na série). Perceba que o FAC da esquerda apresenta a maioria das colunas acima do limite crítico de significância, ou seja, as observações apresentam "memória" do que aconteceu no passado. Perceba também que o gráfico de autocorrelação parcial (FACP) apresenta coeficientes acima do limite crítico em períodos de tempo cíclicos, indicando claramente a existência de sazonalidade na série original, o que impede a utilização da série para previsões. Os gráficos do lado direito mostram os coeficientes de autocorrelação da série já transformada para retirada das influências cíclicas e da memória do valor medido no tempo anterior. Todas as colunas ficam abaixo do limite crítico indicado pelas linhas azuis nas posições de +1,96 e -1,96, unidades de distância do zero. Esses limites se aplicam para os casos em que o *p-valor* é de 0,050.

Gráfico 10.6. Comparação entre FAC e FACP de dois modelos





As transformações realizadas na série original para torna-la estacionária foram realizadas a partir do método ARIMA. Nos anos 1970, Box e Jenkis introduziram uma nova forma de analisar séries temporais estacionárias para previsões. Chama-se método ARIMA.

10.4 Teste Autoregressivo com Médias Móveis Integradas (ARIMA)

Para realizar o teste ARIMA, o primeiro passo é transformar qualquer série de observações em uma série estacionária. Uma vez estabilizada a série, são calculadas as funções de autocorrelação simples e parcial para, através dos padrões dos gráficos, identificar qual a função de autocorrelação mais adequada para a série. Em seguida, são estimados os coeficientes e analisados os resíduos, que são a diferença entre os valores observados e os teóricos, para comprovar a adequação do ajuste do modelo.

Com o modelo válido definido, tem-se que cada valor observado é influenciado por valores em momentos anteriores e expressa uma relação linear em função dos:

- valores recentes da variável;
- ruídos dos valores recentes da variável;
- valores mais distantes da variável;

- ruídos dos valores mais distantes da variável.

Para considerar essas relações, são combinados três termos no modelo ARIMA:

AR, representado pela letra (p), indica o termo de processo autoregressivo. É responsável pela modelagem das influências de valores anteriores da série (Xt-1), ou seja, os efeitos de curto prazo.

MA, representado pela letra (q), indica o resultado do processo de médias móveis. É responsável pela modelagem dos valores dos ruídos da série (Zt-1), ou seja, identifica os efeitos de longo prazo.

I, representado pela letra (d), corresponde ao processo final de integração entre os dois fatores anteriores, dos efeitos de curto e longo prazo.

Assim, o modelo ARIMA (p,d,q) é representado pelo número de retornos autoregressivos, de integração dos valores e do número de período de tempo para a tomada da média móvel. Um modelo ARIMA (1,0,1) padrão indica um retorno autoregressivo e uma unidade de tempo para médias móveis. Como exemplo, vamos aplicar o teste ARIMA (1,0,1) para a variável taxa de desemprego brasileira entre 2003 e 2012. O objetivo é verificar apenas quais são os efeitos de curso prazo (AR) e de longo prazo (MA) sobre os valores da sequência da série. Os resultados obtidos pelo programa Gretl são os que seguem:

Quadro 10.1. Teste ARIMA (1,0,1) para a taxa de desemprego brasileira

Modelo: ARMA, usando as observações 2003:01-2012:12 (T = 120)
Variável dependente: Desemprego
Erros padrão baseados na hessiana

	Coeficiente	Erro Padrão	z	p-valor
const	8,33021	1,82941	4,553	<0,0001****
phi_1	0,981130	0,0168989	58,06	<0,0001****
theta_1	0,0815047	0,0925403	0,8807	0,3785

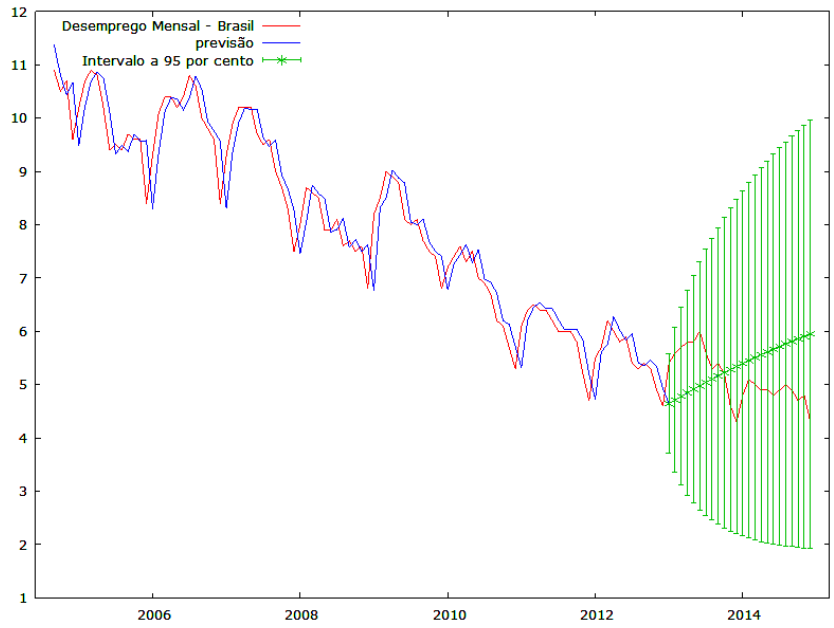
Média var. dependente	8,725833	D.P. var. dependente	2,282984
Média de inovações	−0,039403	D.P. das inovações	0,475280
Log da verossimilhança	−82,73393	Critério de Akaike	173,4679
Critério de Schwarz	184,6178	Critério Hannan-Quinn	177,9959

	Real	Imaginária	Módulo	Frequência
AR				
Raiz 1	1,0192	0,0000	1,0192	0,0000
MA				
Raiz 1	-12,2692	0,0000	12,2692	0,5000

O gráfico a seguir indica os valores da previsão da taxa de inflação para os anos seguintes à série, 2013 e 2014. Isso é feito para podermos comparar com os percentuais reais. Perceba que, como não corrigimos o problema da estacionaridade, o intervalo de

confiança para as previsões é muito grande, variando de 2% a 9% apenas 24 meses depois do fim da série conhecida. A tabela ao lado do gráfico mostra os valores reais do desemprego, os previstos e os intervalos de confiança a 95%. Perceba que os intervalos são muito grandes. Para corrigir o problema, é preciso tornar a série estacionária.

Gráfico 10.7 – Série de previsão de modelo não ajustado



Para intervalos de confiança de 95%, z(0,025) = 1,96.

Obs.	Desemprego	Previsão	Erro padrão	Intervalo a 95%
2013:01	5,40	4,64	0,48	(3,71, 5,57)
2013:02	5,60	4,71	0,69	(3,35, 6,07)
2013:03	5,70	4,78	0,85	(3,11, 6,45)
2013:04	5,80	4,85	0,98	(2,92, 6,77)
2013:05	5,80	4,91	1,09	(2,77, 7,05)
2013:06	6,00	4,98	1,19	(2,65, 7,30)
2013:07	5,60	5,04	1,27	(2,54, 7,53)
2013:08	5,30	5,10	1,35	(2,45, 7,75)
2013:09	5,40	5,16	1,42	(2,38, 7,94)
2013:10	5,20	5,22	1,49	(2,31, 8,13)
2013:11	4,60	5,28	1,55	(2,25, 8,31)
2013:12	4,30	5,34	1,60	(2,20, 8,47)
2014:01	4,80	5,40	1,65	(2,15, 8,63)
2014:02	5,10	5,45	1,70	(2,11, 8,78)
2014:03	5,00	5,51	1,75	(2,08, 8,92)
2014:04	4,90	5,56	1,79	(2,05, 9,06)
2014:05	4,90	5,61	1,83	(2,03, 9,19)
2014:06	4,80	5,66	1,87	(2,00, 9,31)
2014:07	4,90	5,71	1,90	(1,98, 9,43)
2014:08	5,00	5,76	1,93	(1,97, 9,55)
2014:09	4,90	5,81	1,97	(1,95, 9,66)
2014:10	4,70	5,86	2,00	(1,94, 9,77)
2014:11	4,80	5,91	2,02	(1,93, 9,87)
2014:12	4,30	5,95	2,05	(1,93, 9,97)

Como já vimos, a série original não é estacionária e a ARIMA (1,0,1) não foi suficiente para acabar com os efeitos de sazonalidade. Isso porque o coeficiente *theta*, que indica as médias móveis, não é estatisticamente significativo. Por outro lado, o coeficiente *Phi*, que indica o coeficiente autorregressivo, ficou abaixo do limite crítico, indicando significância estatística. Outros indicadores de que o modelo não é estacionário podem ser encontrados nos valores baixos dos critérios Akaike, Hannan-Quinn e Schwarz. Por fim, os valores reais e do módulo para AR e MA ficaram muito próximos, o que indica não estacionaridade. Nesse caso, é preciso alterar o modelo até que se encontrem coeficientes que indiquem que a série se tornou estacionária.

O melhor modelo ARIMA para tornar a série estacionária é o que segue no quadro abaixo, representado por (2,0,4). Foram necessários dois retardos autorregressivos (*phi* 1 e 2) e quatro de médias móveis (*theta* de 1 a 4) para que o modelo se torne significativo, ou seja, estacionário. A representação gráfica da autocorrelação deste modelo é a do gráfico 10.8 acima e à direita. Todos os coeficientes individuais são estatisticamente significativos. Além disso, os critérios apresentaram um ganho em relação ao modelo anterior e os valores do AR e MA do módulo ficam distantes dos valores reais nas raízes sazonais.

Quadro 10.2. Teste ARIMA (2,0,4) para a taxa de desemprego brasileira

Modelo: ARMA, usando as observações 2003:01-2012:12 (T = 120)
Variável dependente: Desemprego
Erros padrão baseados na hessiana

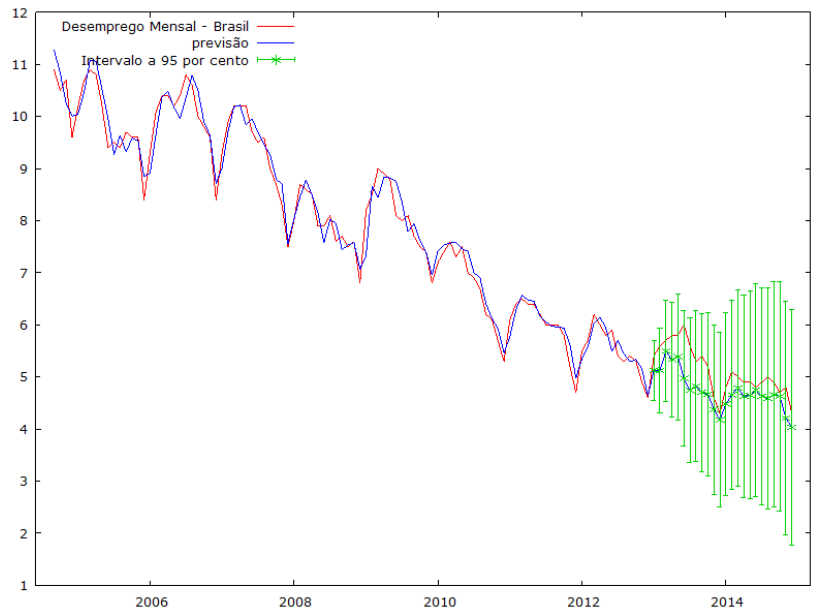
	Coeficiente	Erro Padrão	z	p-valor	
const	8,36624	2,25365	3,712	0,0002	***
phi_1	1,89506	0,162320	11,67	<0,0001	***
phi_2	−0,896570	0,159111	−5,635	<0,0001	***
theta_1	−0,933120	0,143932	−6,483	<0,0001	***
Theta_1	0,210734	0,120287	1,752	0,0798	*
Theta_2	0,420848	0,130118	3,234	0,0012	***
Theta_3	0,707659	0,167605	4,222	<0,0001	***
Theta_4	0,459357	0,130829	3,511	0,0004	***

Média var. dependente	8,725833	D.P. var. dependente	2,282984
Média de inovações	−0,029180	D.P. das inovações	0,296111
Log da verossimilhança	−43,89001	Critério de Akaike	105,7800
Critério de Schwarz	130,8674	Critério Hannan-Quinn	115,9681

		Real	Imaginária	Módulo	Frequência
AR					
	Raiz 1	1,0175	0,0000	1,0175	0,0000
	Raiz 2	1,0961	0,0000	1,0961	0,0000
MA					
	Raiz 1	1,0717	0,0000	1,0717	0,0000
MA (sazonal)					
	Raiz 1	-1,2298	-0,8151	1,4755	-0,4068
	Raiz 2	-1,2298	0,8151	1,4755	0,4068
	Raiz 3	0,4596	-0,8881	1,0000	-0,1740
	Raiz 4	0,4596	0,8881	1,0000	0,1740

Agora, com o modelo ajustado, podemos fazer as previsões para o período posterior a 2012. Vamos manter previsões para a inflação nos 24 meses seguintes, indo até 2014. Os resultados são os que seguem:

Gráfico 10.8 – Série de previsão de modelo não ajustado



Para intervalos de confiança de 95%, z(0,025) = 1,96.

Obs.	Desemprego	Previsão	Erro padrão	Intervalo a 95%
2013:01	5,40	5,13	0,30	(4,54, 5,70)
2013:02	5,60	5,12	0,41	(4,31, 5,92)
2013:03	5,70	5,50	0,49	(4,52, 6,46)
2013:04	5,80	5,34	0,56	(4,23, 6,43)
2013:05	5,80	5,39	0,62	(4,18, 6,59)
2013:06	6,00	4,97	0,66	(3,67, 6,27)
2013:07	5,60	4,74	0,70	(3,36, 6,12)
2013:08	5,30	4,82	0,74	(3,37, 6,27)
2013:09	5,40	4,71	0,77	(3,18, 6,22)
2013:10	5,20	4,67	0,80	(3,09, 6,24)
2013:11	4,60	4,38	0,83	(2,74, 6,00)
2013:12	4,30	4,18	0,86	(2,49, 5,85)
2014:01	4,80	4,48	0,89	(2,72, 6,23)
2014:02	5,10	4,66	0,93	(2,83, 6,47)
2014:03	5,00	4,79	0,96	(2,90, 6,67)
2014:04	4,90	4,63	0,99	(2,68, 6,56)
2014:05	4,90	4,65	1,02	(2,65, 6,64)
2014:06	4,80	4,75	1,04	(2,70, 6,79)
2014:07	4,90	4,63	1,07	(2,54, 6,71)
2014:08	5,00	4,59	1,09	(2,45, 6,71)
2014:09	4,90	4,66	1,11	(2,49, 6,83)
2014:10	4,70	4,63	1,12	(2,42, 6,83)
2014:11	4,80	4,21	1,14	(1,97, 6,44)
2014:12	4,30	4,03	1,16	(1,76, 6,30)

Perceba que, agora, todos os valores previstos (coluna previsão da tabela ao lado do gráfico) ficam muito próximos das taxas reais de desemprego em 2013 e 2014, além disso, os intervalos de confiança incluem todos os valores reais do período da previsão. Isso indica que se não conhecêssemos os valores reais, nossa variável preditiva estaria bem ajustada.

Como vimos nos exemplos acima, as partes AR e MA de uma série temporal são componentes irregulares que podem ser modelados pelo método de Box e Jenkins para extrair as partes relevantes das séries, usando o que elas têm de substantivo para predizer valores futuros.

10.5 Teste para Raízes Unitárias

Além dos elementos de oscilação de curto e longo prazo, uma série temporal também pode apresentar oscilações randômicas. Esse tipo de oscilação é chamada de raiz unitária. A presença de raiz unitária em uma série indica que ela não é estacionária. Quando há um impacto sobre os dados da série e seus efeitos são permanentes – ainda que randômicos – as mudanças nos valores futuros tornam-se uma condição necessária. Esses impactos em determinados pontos da série não desaparecem no tempo, são integrados ou absorvidos pela série, gerando efeitos contínuos (Pevehouse & Prozek, 2008). O teste mais usado para verificar a existência de raízes unitárias em uma série temporal é o Teste de Dickey-Fuller Aumentado (ADF). Neste teste, a hipótese nula (H0) é que todas as oscilações identificadas na série são randômicas e a série é estacionária. Se o *p-valor* ficar acima do limite crítico, significa que há uma chance superior à aceitável de erro se considerarmos que só há variações randômicas na série. Neste caso, é preciso considerar a existência de raiz unitária na série e ela apresentar algum tipo de tendência ou efeito cíclico. Os casos em que a hipótese nula não pode ser rejeitada são os que a série apresenta raiz unitária e cujas variações não podem ser consideradas aleatórias (Pevehouse & Prozek, 2008).

Os resultados dos testes DF são comparados a valores tabelados para identificação dos limites críticos de significância estatística. Para número de casos ao redor de 100, os valores tabelados para limites críticos são os seguintes: *p-valor* 0,001 = -2,60;

p-valor 0,050 = -1,95; e *p-valor* 0,100 = -1,61. Rejeitamos a hipótese nula quando o valor encontrado ficar abaixo dos limites críticos tabelados. O teste Dickey-Fuller Aumentado é indicado para identificação de raízes unitárias e para presença de tendência determinista. Para exemplificar o teste de raiz unitária completo, segue abaixo o ADF para a variável taxa de desemprego entre 2003 e 2012 no Brasil, pelo Gretl.

Quadro 10.3. Teste ADF para a taxa de desemprego brasileira

Teste Aumentado de Dickey-Fuller para Desemprego
Testar para baixo a partir de d 12efasagens, critério AIC
Tamanho da amostra: 120
Hipótese nula de raiz unitária: a = 1

Teste sem constante
Incluindo 12 defasagens de (1-L) Desemprego
modelo: (1-L)y = (a-1)*y(-1) + ... + e
Valor estimado de (a - 1): -0,0108068
Estatística de teste: tau_nc(1) = -2,3947
p-valor assintótico 0,01609
Coeficiente de 1ª ordem para e: -0,022
Diferenças defasadas: F(12, 107) = 4,905 [0,0000]

Teste com constante
Incluindo 12 defasagens de (1-L) Desemprego
modelo: (1-L)y = b0 + (a-1)*y(-1) + ... + e
Valor estimado de (a - 1): -0,00606879
Estatística de teste: tau_c(1) = -0,334604
p-valor assintótico 0,9175
Coeficiente de 1ª ordem para e: -0,019
Diferenças defasadas: F(12, 106) = 4,860 [0,0000]

Com constante e tendência
Incluindo 12 defasagens de (1-L) Desemprego
modelo: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
Valor estimado de (a - 1): -0,280506
Estatística de teste: tau_ct(1) = -3,025
p-valor assintótico 0,1252
Coeficiente de 1ª ordem para e: -0,050
Diferenças defasadas: F(12, 105) = 3,611 [0,0002]

O *software* Gretl calcula o número de defasagens automaticamente em função do N da série. No caso, ele propõe 12 defasagens e utiliza o critério AIC. Os resultados incluem uma série de estatísticas em três testes distintos. O primeiro sem constante, o segundo apenas com constante e o terceiro com constante e tendência. Perceba que eles apresentam valores distintos, mas em todos os casos o *p-valor* está acima de 0,050, indicando a existência de raiz unitária em todos os casos. Ou seja, a série de desemprego não é estacionária em nenhum dos casos, necessitando de transformações para ser usada na geração de coeficientes de predição.

10.6 Análise multivariada no tempo (efeitos de intervenção e de transferência)

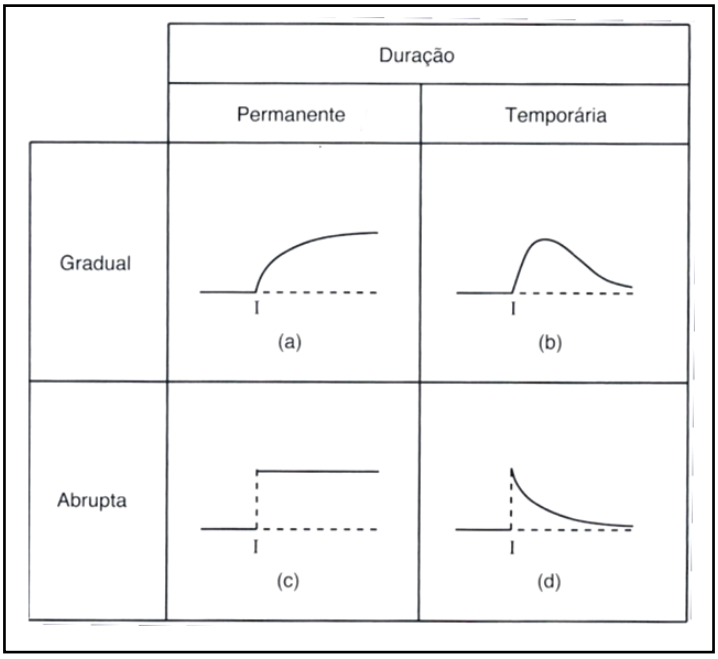
Além de analisar as variações de uma série usando a passagem do tempo como explicação, é possível considerar o efeito de uma variável explicativa externa, que não o tempo, nas variações da série que estamos estudando. Por exemplo, imagine uma série com unidade mensal do número de mortos em acidentes de trânsito no Brasil entre os anos de 1980 e 2015. Espera-se uma tendência crescente na série, entre outros motivos, pelo aumento no número de veículos nas rodovias e do País. No entanto, em 1997 houve um evento externo, que foi a entrada em vigor do Código Brasileiro de Trânsito (lei 9.503/97), que estabeleceu formas mais claras de fiscalização e, principalmente, endureceu multas e punições para infratores. É possível considerar que, a partir da entrada em vigor da lei, esse evento externo apresente alguma intervenção sobre a série histórica de mortos em acidentes de trânsito no Brasil. Por intervenção entende-se a ocorrência de um tipo de evento em dado momento do tempo T, que pode manifestar-se a partir de então de forma permanente ou temporária na série (Morettin & Toloi, 2004). A análise de intervenção visa medir o impacto de um evento sobre o comportamento da série. Em geral, esse impacto pode ser como uma (i) função degrau, quando a intervenção é permanente; ou (ii) função impulso, quando o efeito é temporário. A intervenção se dá sempre por uma variável *dummy*, por identificar os dois períodos de tempo, com e sem a característica, antes e depois da intervenção.

Se pretendermos medir a intervenção de uma variável explicativa contínua sobre as alterações de uma série temporal, então, trata-se de um efeito de transferência. Seguindo o caso anterior, poderíamos acrescentar como variável de transferência o número de veículos que circulam no país a cada período do tempo para controlar o efeito da passagem temporal sobre o número de acidentes. Com isso, a variação no número de veículos, que deve apresentar tendência similar à do número de acidentes, ajudará a controlar o efeito da passagem do tempo.

Podemos definir intervenção ou transferência de um efeito como a ocorrência em determinado momento do tempo, conhecido *a priori*, e que se manifesta por um in-

tervalo de tempo posterior, afetando temporária ou permanentemente a série em análise. Em geral, o efeito de uma intervenção muda o nível da série ou sua inclinação. No entanto, é preciso considerar que existem três fontes de ruídos que podem deixar o efeito da intervenção invisível: tendência, sazonalidade e o erro aleatório (Morettim & Tolloi, 2004). Vale lembrar ainda que existe a possibilidade do impacto do evento na curva ser nulo. Porém, quando existem efeitos, a diferença entre os padrões está relacionada com a duração da mudança, que pode ser permanente ou temporária, e com a magnitude (também chamada de formato), que pode ser abrupta ou gradual. Sendo assim, a mudança causada por um componente interveniente da opinião pode ser permanente e gradual ou permanente e abrupta; se a mudança for temporária, ela também pode ser gradual ou abrupta, conforme indica a figura a seguir.

Figura 10.1. Formatos de intervenção externa em curvas temporais



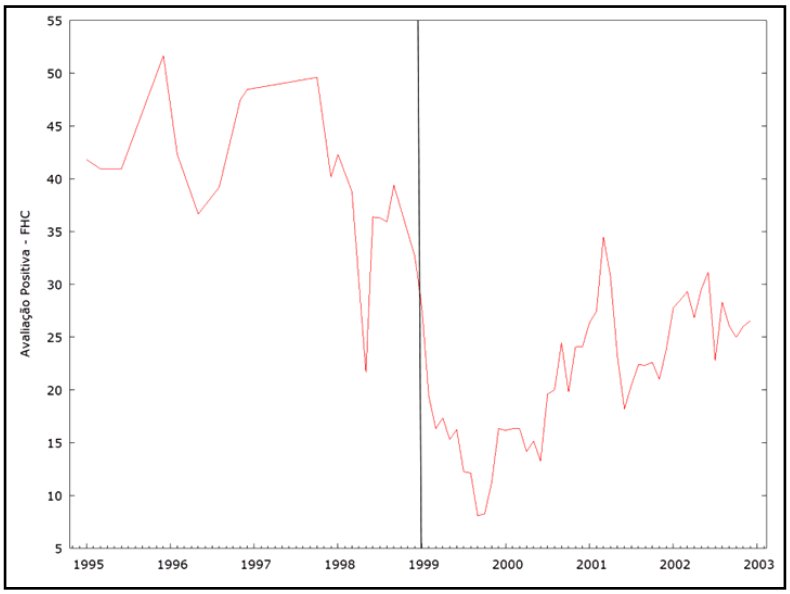
Fonte: Morettin e Toloi, 2004, p. 285.

Três desses modelos de impacto podem ser determinados por um componente interveniente simples (a, c e d). O padrão b (impacto gradual e temporário) não pode ser identificado tão facilmente. Esse padrão costuma ser menos útil entre os quatro por conta das diferentes variáveis intervenientes que podem atuar nele (Got-

man, 1984).

Para testar o efeito de intervenção a partir do *software* Gretl, vamos usar duas novas variáveis aqui. A primeira é a variação da avaliação positiva do governo Fernando Henrique Cardoso (FHC) durante o período de 1995 a 2002. Esse é a variável dependente. Consideramos que a passagem do tempo explica parte das variações, no entanto, é possível que a mudança do primeiro para o segundo mandato apresente uma intervenção para além da simples passagem de tempo. Então, usaremos uma variável *dummy* com os códigos 1 = primeiro mandato e 0 = segundo mandato. Com isso, pretendemos medir a mudança da avaliação positiva do governo FHC ao longo do tempo, considerando a intervenção do segundo mandato. O gráfico da série temporal da avaliação positiva do governo FHC é o que segue.

Gráfico 10.9. Série temporal Avaliação Positiva governo FHC



Percebe-se que a série apresenta uma tendência, mas também uma mudança abrupta justamente no momento de mudança de primeiro para segundo mandato, em 1999. Também existem oscilações cíclicas. Para rodar o teste Autoregressivo de ordem 1 (AR1) no Gretl, coloca-se a variável avaliação positiva de FHC na caixa de variável dependente e, na caixa de variável independente, a "constante" e a variável *dummy* "primeiro mandato". Aqui, utiliza-se o método de Prains-Winsten. Os resultados constam no quadro abaixo.

Quadro 10.4. Teste AR1 na avaliação FHC

Modelo: Prais-Winsten, usando as observações 1995:01-2002:12 (T = 96)
Variável dependente: Avaliação Positiva
rho = 0,932242

	Coeficiente	Erro Padrão	razão-t	p-valor	
Constante	28,9902	4,81559	6,020	<0,0001	***
É Primeiro Mandato	**6,78281**	3,28884	2,062	0,0419	**

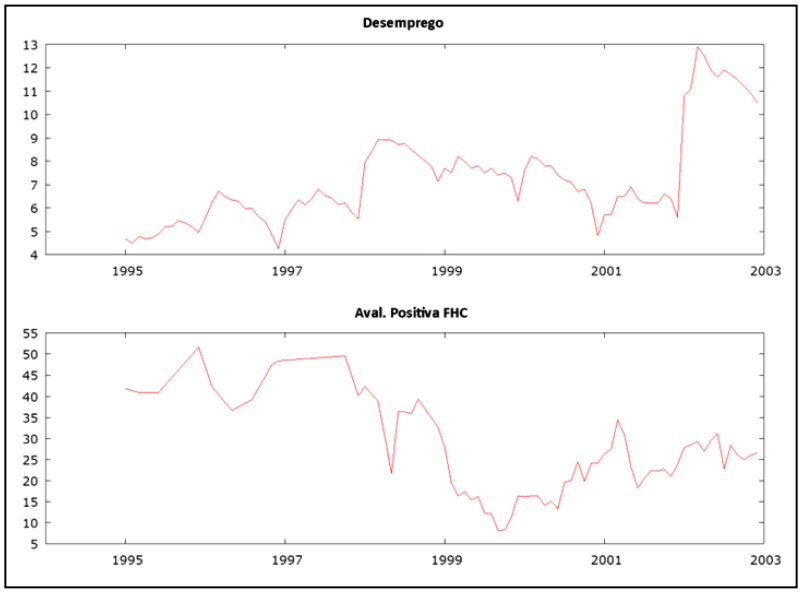
Estatísticas baseadas nos dados rô-diferenciados:

Média var. dependente	31,86750	D.P. var. dependente	12,25276
Soma resíd. quadrados	1092,003	E.P. da regressão	3,408379
R-quadrado	**0,923602**	R-quadrado ajustado	0,922789
F(1, 94)	17,63296	P-valor(F)	0,000061
Rô	0,009034	Durbin-Watson	**1,977451**

Os coeficientes individuais mostram que a variável *dummy* "primeiro mandato" apresenta um efeito estatisticamente significativo sobre a avaliação positiva do governo FHC. O coeficiente de 6,78 e *p-valor* 0,0001 indicam que no primeiro mandato o governo FHC teve avaliação positiva superior ao segundo mandato. Além disso, as estatísticas do modelo mostram um bom ajustamento. A média da variável dependente é a média de avaliação positiva ao longo dos dois mandatos (31,86%). O r² de 0,923 indica alto ajustamento do modelo explicativo. Além disso, o coeficiente *Durbin-Watson* (1,977) fica entre 1,85 e 2,15, que é o intervalo de valores que indica que não há autocorrelação entre as variáveis do modelo.

Agora, para apresentar um efeito de transferência de uma variável explicativa contínua para uma série temporal, manteremos como variável dependente a série de avaliação positiva do período FHC. A variável independente/de transferência será a taxa de desemprego mensal no período de 1995 a 2002. Todos os demais critérios do modelo AR(1) serão mantidos. Os resultados estão nos gráficos a seguir.

Gráfico 10.10. Comparação entre séries temporais entre 1995 e 2002



Os gráficos acima mostram que, enquanto os percentuais de desemprego variaram em torno de 6% entre 1995 e 2002, a série apresenta uma ruptura com salto para algo em torno de 10% no último ano do período. A série de avaliação positiva de FHC, já conhecida, mostra uma queda em 1999, sem grandes oscilações no último ano do período. Como pretendemos medir o impacto de transferência da variação da taxa de desemprego sobre a avaliação de governo, não podemos considerar as duas medições no mesmo ponto do tempo. É preciso dar um retorno (t-1) para a variável "Desemprego". Assim, o modelo construído irá medir o impacto do desemprego no mês anterior sobre a avaliação positiva do governo FHC. Por isso, no modelo, a variável desemprego aparece com a notação "Desemprego_1".

Os resultados completos estão nos quadros a seguir. É possível perceber que no modelo autoregressivo, aquele que retira parte da "memória" do passado na variável dependente, mostra que o coeficiente não é estatisticamente significativo. Ainda que negativo (-0,557), ele não apresenta significância estatística, o que indica que as mudanças nas taxas de desemprego não apresentam impacto sobre a variação das avaliações de governo. A estatística Durbin-Wattson acima de 1,85 mostra pouco efeito de autocorrelação entre os valores. O r^2 em 0,922 mostra que o modelo temporal é bem ajustado, ainda que sem o impacto das mudanças nas taxas de desemprego para a avaliação positiva do governo.

Quadro 10.5. Modelos de regressão para desemprego sobre avaliação FHC

Modelo: Prais-Winsten, usando as observações 1995:01-2002:12 (T = 96)
Variável dependente: Avaliação Positiva
rho = 0,956114

	Coeficiente	Erro Padrão	razão-t	p-valor	
Constante	36,6068	7,44822	4,915	<0,0001	***
Desemprego_1	**−0,557482**	0,470220	−1,186	0,2388	

Estatísticas baseadas nos dados rô-diferenciados:

Média var. dependente	31,86750	D.P. var. dependente	12,25276
Soma resíd. quadrados	1111,438	E.P. da regressão	3,438577
R-quadrado	**0,922191**	R-quadrado ajustado	0,921364
F(1, 94)	10,64739	P-valor(F)	0,001537
Rô	0,061941	Durbin-Watson	**1,871955**

Modelo: MQO, usando as observações 1995:01-2002:12 (T = 96)
Variável dependente: Avaliação Positiva

	Coeficiente	Erro Padrão	razão-t	p-valor	
Constante	50,1138	4,24695	11,80	<0,0001	***
Desemprego_1	**−2,55226**	0,572173	−4,461	<0,0001	***

Média var. dependente	31,86750	D.P. var. dependente	12,25276
Soma resíd. quadrados	11770,78	E.P. da regressão	11,19022
R-quadrado	**0,174696**	R-quadrado ajustado	0,165916
F(1, 94)	19,89739	P-valor(F)	0,000023
Log da verossimilhança	−367,0514	Critério de Akaike	738,1029
Critério de Schwarz	743,2316	Critério Hannan-Quinn	740,1760
Rô	0,942799	Durbin-Watson	**0,115595**

Para comparar diferentes regressões, a segunda parte do quadro acima mostra os resultados obtidos com outro modelo de regressão, pelo método dos Mínimos Quadrados Ordinários (MQO), que desconsidera os efeitos de autocorrelação de valores no tempo. Perceba que o coeficiente da variável desemprego é estatisticamente significativo, com *p-valor* 0,001 e coeficiente -2,552 – o que indicaria que o crescimento nos índices de desemprego geraram uma queda na avaliação positiva do governo FHC. No entanto, as estatísticas do modelo mostram um baixo ajustamento, com r² de apenas 0,174 e estatística *Durbin-Watson* de 0,115. Portanto, muito abaixo do limite crítico para ausência de autocorrelação de valores. Ou seja, o modelo de regressão tradicional MQO dá a impressão que a variável explicativa apresenta efeito significativo, porém em um modelo pouco explicativo.

Para concluir o exemplo, vamos rodar uma regressão AR(1) com os dois tipos de efeitos, intervenção e transferência, sobre a série temporal da avaliação positiva do governo FHC. Assim, ainda no modelo Prains-Winsten, a variável dependente continua

sendo Avaliação Positiva. Agora, serão duas variáveis explicativas além da constante: a variável de intervenção, “primeiro mandato”, e a variável de transferência, “taxa de desemprego” com um retardo no tempo (desemprego_1). Os resultados são apresentados no quadro a seguir.

Quadro 10.6. Teste AR(1) com dois efeitos sobre avaliação FHC

Modelo: Prais-Winsten, usando as observações 1995:01-2002:12 (T = 96)
Variável dependente: Avaliação Positiva
rho = 0,925982

	Coeficiente	Erro Padrão	razão-t	p-valor	
Constante	33,3894	5,55778	6,008	<0,0001	***
Desemprego_1	**−0,642969**	0,462223	−1,391	0,1675	
É Primeiro Mandato	**7,15485**	3,24720	2,203	0,0300	**

Estatísticas baseadas nos dados rô-diferenciados:

Média var. dependente	31,86750	D.P. var. dependente	12,25276
Soma resíd. quadrados	1070,386	E.P. da regressão	3,392569
R-quadrado	**0,925147**	R-quadrado ajustado	0,923537
F(2, 93)	10,58265	P-valor(F)	0,000072
rô	0,011464	Durbin-Watson	**1,975145**

Os coeficientes individuais mostram que a variável “desemprego”, com um retorno ainda que com coeficiente negativo, não apresenta significância estatística sobre a variação de avaliação positiva de governo. Por outro lado, a variável de intervenção “primeiro mandato” tem efeito estatisticamente significativo para avaliação de governo. O coeficiente positivo de 7,15 significa que no primeiro mandato a avaliação positiva do governo FHC foi significativamente superior aos percentuais do segundo mandato. O modelo mostra-se bem ajustado ($r^2 = 0,925$) e sem efeito de autocorrelação entre as observações, com *Durbin-Watson* acima de 1,85. A regressão AR(1) com os dois tipos de variáveis independentes mostra que, na explicação das variações da série temporal de avaliação positiva do governo FHC, o efeito de intervenção (primeiro governo) tem mais impacto que o efeito de transferência (taxa de desemprego). Se você está achando estranho esse resultado, por considerar que desemprego teria um impacto lógico sobre a avaliação de um governo, sugiro que não deixe de fazer os exercícios propostos neste capítulo. Os resultados dele ajudam a entender a dinâmica de impacto das variáveis explicativas sobre a avaliação do governo FHC.

Neste capítulo, foram apresentados os princípios básicos, características e as

principais técnicas para análises de séries temporais. O que caracteriza essas séries é, ao contrário das regressões discutidas em outros capítulos, considerar a passagem do tempo como a variável explicativa. Portanto, análise de séries temporais é um tipo específico de regressão que leva em conta a passagem do tempo. Vimos que essa característica básica tem uma implicação fundamental para os testes, que é a quebra do pressuposto da independência entre as observações. Se o tempo é uma variável, então a ocorrência de um fenômeno em um momento específico carrega uma "memória" do que aconteceu no passado e influenciará, pelo menos em parte, o que virá a seguir. Para corrigir a quebra do pressuposto da independência de observações nas séries temporais, foram apresentadas técnicas para identificação de dados autocorrelacionados e série com raiz unitária. Também foi apresentada a principal técnica de predição a partir de dados já existentes, a ARIMA. Por fim, discutimos rapidamente efeitos de variáveis externas, independentes, no impacto do tempo sobre a variável dependente. Os impactos podem ser graduais ou abruptos, permanentes ou temporários nas mudanças da série temporal em análise.

10.7 Referências bibliográficas do Capítulo X

Box, G., & Jenkins, G. (1976). *Time Series Analysis: forecasting and control*. San Francisco: Holden-Dan.

Gotman, J. M. (1984). *Time-Series analysis: a comprehensive introduction for social scientists*. Cambridge: Cambridge University Press.

Gujarat, D. (2006). *Econometria Básica*. Rio de Janeiro: Ed. Elsevier.

Morettin, P., & Toloi, C. (2004). *Análise de Séries Temporais*. São Paulo: Editora Edgard Blücher.

Pevehouse, J. C., & Brozek, J. D. (2008). Time-series *analysis*. In J. Box-Steffensmeier, H. E. Brady, & D. Collier (Ed.). *The Oxford Handbook of Political Methodology,* 10, 456-471. New York: Oxford University Press.

10.8 Exercícios propostos do Capítulo X

10.8.1 Utilizando o banco de dados BDCAP11V2_TimeSeries disponível em link, considere as seguintes variáveis:
Dependente = Avaliação negativa do governo FHC;
Explicativas =(i) Taxa de desemprego entre 1995 e 2002 e (ii) Se está ou não no Primeiro Mandato.

No *software* Gretl insira o banco de dados e faça o seguinte:

10.8.1 Gere o gráfico de médias móveis simples (centradas em 12 meses) para a variável dependente;

10.8.2 Gere e analise os gráficos de autocorreção FAC e FACP, propondo ajustes necessários ao modelo;

10.8.3 Rode o teste de Raízes Unitárias ADF para a avaliação negativa de FHC e responda se é possível considerar a série como estacionária ou não.

10.8.4 Ajuste o melhor modelo ARIMA (p,d,q), começando por (1,0,1) para a variável dependente e depois ajustando até a retirada do efeito de memória. Rode um correlograma do modelo para testar a existência de memória.

10.8.5 Calcule um modelo autoregressivo (AR1) para descrição das relações ao longo do tempo entre Avaliação negativa de FHC, como dependente, e “retorno da taxa de desemprego” e “se está ou não no Primeiro Mandato” como explicativas. Analise os resultados.

ANEXO DO CAPÍTULO X

ANEXO 10.1 – PASSO A PASSO PARA INSERÇÃO DE DADOS NO GRETL

O arquivo para instalação do Gretl pode ser baixado a partir do endereço eletrônico: http://gretl.sourceforge.net/win32/index_pt.html, selecionando posteriormente o sistema operacional em que o programa será instalado.
A partir daí, faz-se o download do arquivo executável "gretl-2018.exe" até a instalação completa do programa.

Uma vez instalado o programa, abre-se a janela do Gretl para importar o banco de dados com o qual se trabalhará.

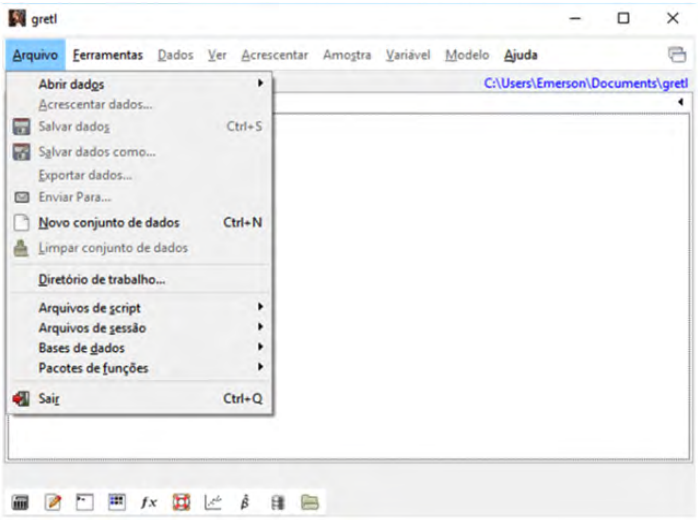
Seguir o caminho:

Arquivo /

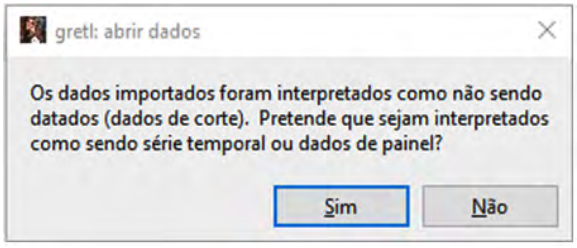
Abrir Dados /

Selecionar Arquivos do Usuário

Buscar o arquivo "BDCap11V2_TimeSeries" no local em que ele foi salvo no seu computador.

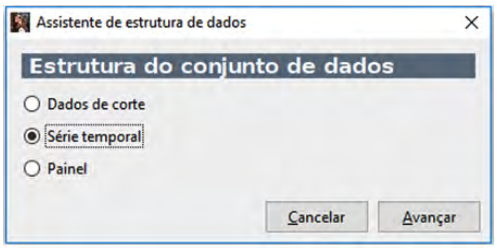


Após importar o banco de dados, serão abertas uma série de janelas para categorização do banco de dados. A primeira pergunta é se os dados são de uma série temporal ou painel? A escolha é SIM.



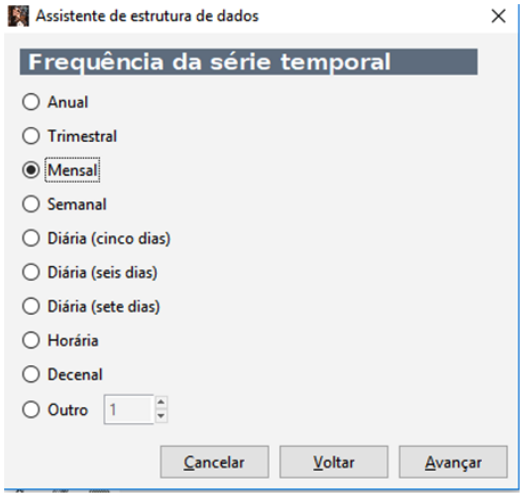
Em seguida, abre-se uma caixa de diálogo para que seja selecionado o tipo de dados. São três opções: dados de corte, série temporal ou painel.

Aqui, selecionamos SÉRIE TEMPORAL.



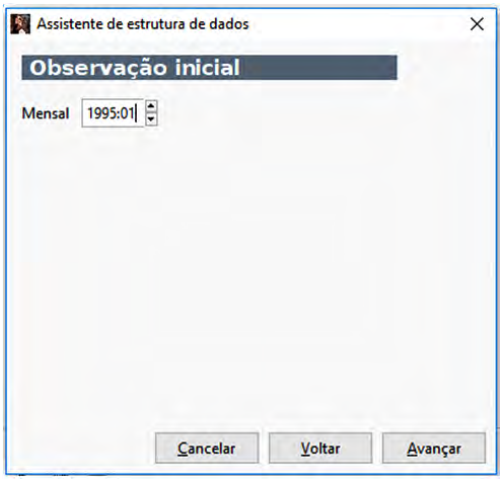
Então, é preciso selecionar qual o tipo de frequência da série temporal. As opções vão de anual a decenal. No nosso caso, o banco de dados está organizado a partir da unidade de medida mensal.

Selecionamos MENSAL.

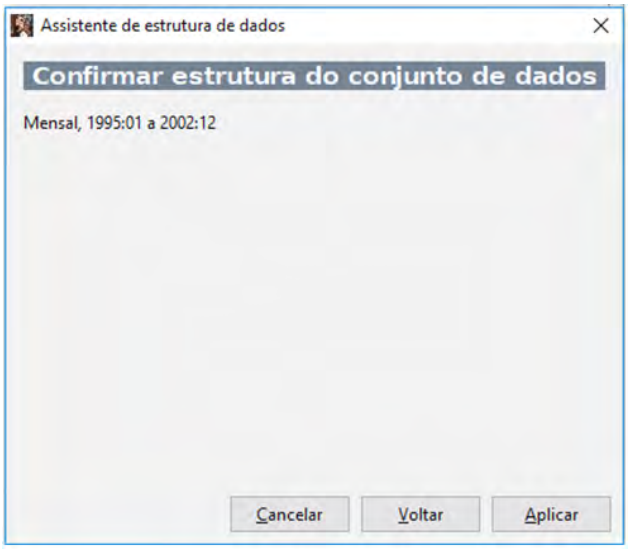


Uma vez selecionada a frequência mensal, é preciso indicar em que ponto do tempo a série começa. No nosso caso, o banco de dados começa em janeiro de 1995.

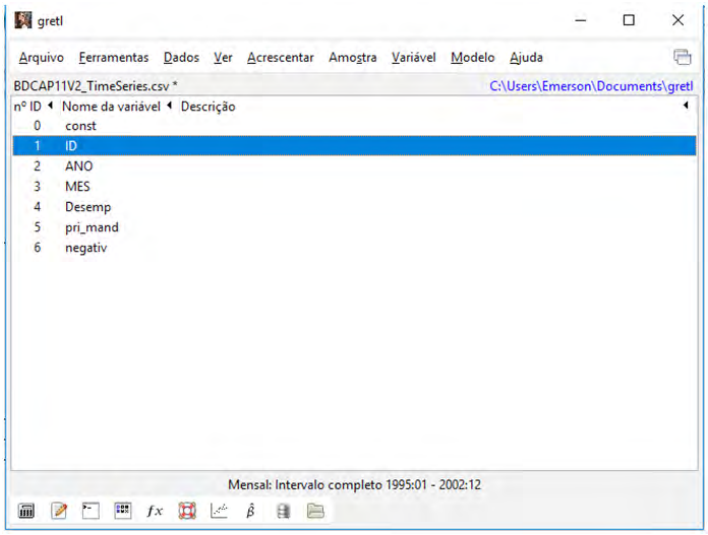
Então, digita-se no formato= 1995:01.

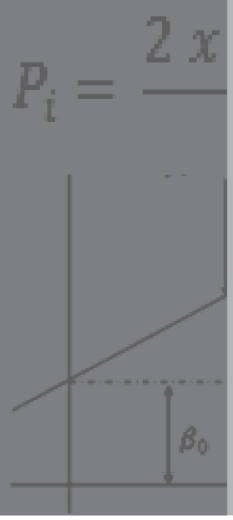


Antes de abrir o programa, o Gretl confirma, em uma tela final, se a série carregada no programa está correta. Ele indica o intervalo temporal a partir da observação inicial indicada.
No nosso caso, o banco de dados vai de janeiro de 1995, até dezembro de 2002.
Como o intervalo está correto, escolha APLICAR.



O banco de dados “BDCap11V2_TimeSeries” possui seis variáveis:
ID = identificador da linha.
ANO = identificador do ano (de 1995 a 2002).
MÊS = identificador do mês (de 1 a a12).
DESEMP = taxa de desemprego mensal em % (IBGE).
PRI_MAND = indica se é primeiro mandato ou não (1 = primeiro mandato, 0 = não é primeiro mandato).
NEGATIV = avaliação negativa (ruim + péssimo) do governo Fernando Henrique Cardoso.
Para gerar gráficos, correlogramas ou estatísticas descritivas, basta clicar com botão direito sobre a variável ou escolher a opção “variável”.
Para rodar testes de regressão ARIMA, Raiz Unitária, AR (1) e outros, escolher opção Modelo/





ADENDO I

GABARITO DE RESPOSTAS AOS EXERCÍCIOS PROPOSTOS

RESPOSTAS ÀS QUESTÕES DO CAPÍTULO I

1.5.1 V de Cramer:

$$v = \sqrt{\frac{\chi^2}{N.(k-1)}} = \sqrt{\frac{5{,}263}{1849\ x\ (1-1)}} = 0{,}053$$

O coeficiente V de Cramer para associação entre sexo e partido é de 0,053 ou associação de 5,3%.

1.5.2 e 1.5.3 Os valores de Delta e resíduos padronizados estão na tabela abaixo:

SEXO	ESTATÍSTICA	PT	PMDB	PSDB	TOTAL
Homem	N	79	686	581	1346
	Valor Esperado	72,8	672,64	600,57	
	Delta	6,2	13,36	-19,57	
	Delta Probabilidade	0,003	0,007	-0,01	
	Resíduo Padronizado	0,73	0,52	-0,8	
Mulher	N	21	238	244	503
	Valor Esperado	27,2	251,36	224,43	
	Delta	-6,2	-13,36	19,57	
	Delta Probabilidade	-0,003	-0,007	0,01	
	Resíduo Padronizado	-1,19	-0,84	1,31	
TOTAL		100	924	825	1849
x^2 = 5,263 (0,072)					

Interpretação: Tanto o teste de diferenças de médias X^2, quando os resíduos mostram que não há diferenças estatisticamente significativas entre as proporções de homens e mulheres eleitos por PT, PMDB e PSDB para deputado estadual em 2014. Ainda que os resíduos de eleitas pelo PMDB e PT sejam negativos e do PSDB, positivo.

Respostas às Questões do Capítulo II

2.4.1.a) Teste de Independência Q_{xy}

$$Q_{xy} = \frac{(BxC) - (AxD)}{(BxC) + (AxD)} = \frac{12.308.154}{14.146.756} = 0{,}870$$

O coeficiente Qxy entre resultado da eleição e ser candidato em coligação de partidos ou não é de 0,870, ou 87% de associação. Significa que candidatos que estão coligados tendem a estar mais no grupo dos eleitos e essa associação é forte.

2.4.1.b) Cálculo de pares consistentes e inconsistentes na tabela quádrupla

$$P_c = \frac{2\ x\ (B\ x\ C)}{N^2} = \frac{26.454.900}{261.630.625} = 0{,}101$$

$$P_i = \frac{2\ x\ (A\ x\ D)}{N^2} = \frac{1.838.592}{261.630.625} = 0{,}007$$

Há uma proporção de 0,101 de pares consistentes e 0,007 de pares inconsistentes, o que indica que algum grau de associação existe entre as duas variáveis, pois a consistência é maior que a inconsistência.

2.4.1.c) Cálculo do MVE

$$MVE = \frac{Marg.Não\ X\ x\ Marg.Y}{N} = \frac{2.445\ x\ 5.522}{16.175} = 872{,}93$$

O número mínimo de valores esperados em uma casa da tabela é de 872,93, portanto, muito acima do necessário para validar os resultados do teste.

2.4.1.d) Identificação do intervalo de confiança

$$1{,}96\ x\ \sqrt{\frac{\left(1-\left(Q_{xy}{}^2\right)\right)^2\ x\ \frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}{4}} = 1{,}96\ x\ \sqrt{\frac{(1-0{,}756)^2\ x\ \frac{1}{8202}+\frac{1}{5410}+\frac{1}{2445}+\frac{1}{112}}{4}} =$$

$$1{,}96\ x\ \sqrt{\frac{(0{,}244)^2\ x\ 0{,}00012+0{,}00018+\ 0{,}00040+\ 0{,}00892}{4}} = 1{,}96\ x\ \sqrt{\frac{0{,}059\ x\ 0{,}0096}{4}} = 1{,}96\ x\ \sqrt{\frac{0{,}000571}{4}}$$

$$= 1{,}96\ x\ \sqrt{0{,}000142} = 1{,}96\ x\ 0{,}0119 = 0{,}0234$$

Limite Superior: 0,870 + 0,0234 = 0,893

Limite Inferior: 0,870 – 0,0234 = 0,846

2.4.2 Etapas para o cálculo do Q_{xy} com variável Teste:

$$Q_{xy}\,ligado = \frac{[(BT\ x\ CT) + (B\overline{T}\ x\ C\overline{T})] - [(AT\ x\ DT) + (A\overline{T}\ x\ D\overline{T})]}{[(BT\ x\ CT) + (B\overline{T}\ x\ C\overline{T})] + [(AT\ x\ DT) - (A\overline{T}\ x\ D\overline{T})]} = \frac{9.960.844}{9.979.468} = 0{,}998$$

$$Q_{xy}\,diferente = \frac{[(BT\ x\ C\overline{T}) + (B\overline{T}\ x\ CT)] - [(AT\ x\ D\overline{T}) + (A\overline{T}\ x\ DT)]}{[(BT\ x\ C\overline{T}) + (B\overline{T}\ x\ CT)] + [(AT\ x\ D\overline{T}) + (A\overline{T}\ x\ DT)]} = \frac{2.347.310}{2.702.126} = 0{,}868$$

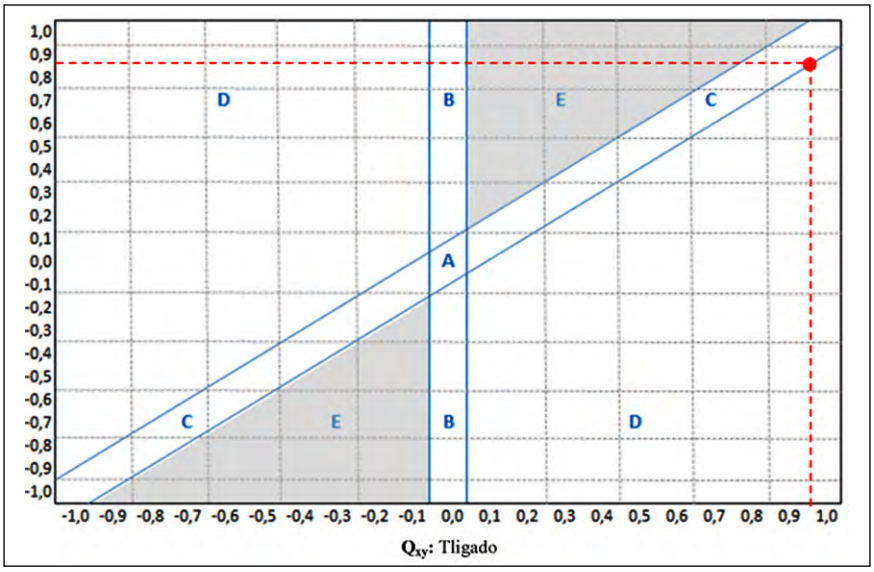
$$P1 = \frac{(BT\ x\ CT) + (B\overline{T}\ x\ C\overline{T}) + (AT\ x\ DT) + (A\overline{T}\ x\ D\overline{T})}{[(BT + B\overline{T})\ x\ (CT + C\overline{T})] + [(AT + A\overline{T})\ x\ (DT + D\overline{T})]} = \frac{11.444.620}{14.146.746} = 0{,}808$$

$$P2 = \frac{(BT\ x\ C\overline{T}) + (B\overline{T}\ x\ CT) + (AT\ x\ D\overline{T}) + (A\overline{T}\ x\ DT)}{[(BT + B\overline{T})\ x\ (CT + C\overline{T})] + [(AT + A\overline{T})\ x\ (DT + D\overline{T})]} = \frac{2.702.126}{14.146.746} = 0{,}191$$

$$Q_{xy:t} = (Q_{xy}\ ligado\ x\ P.ligado) + (Q_{xy}\ diferente\ x\ P.diferente)$$
$$= (0{,}998\ x\ 0{,}808) + (0{,}868\ x\ 0{,}191) = \mathbf{0{,}973}$$

Resultado: A correlação alta entre estar coligado e conseguir se eleger, que inicialmente é de 0,870, sobe ainda mais quando controlada por sexo, chegando a 0,973. Ou seja, estar em partido coligado está associado com conseguir se eleger, mas a associação é ainda mais forte quando se trata de candidata.

O gráfico a seguir mostra que a interseção das retas dos dois coeficientes, antes e depois do Teste, localiza-se no limite entre as áreas C e D, porém, ainda dentro da área C, ou seja, a variável sexo apresenta um efeito de especificação sobre a associação de ordem zero. Isso porque a associação inicial entre partido coligado e sucesso eleitoral já era alta. O que a variável de controle fez, foi especificar que para as mulheres essa associação é ainda maior.



2.4.3 Cálculo do coeficiente Gama

$$G = \frac{PC - PI}{PC + PI} = \frac{25.950.962 - 28.442.890}{25.950.962 + 28.442.890} \frac{-2.491.928}{54.393.852} = -0{,}045$$

	abaixo 10%	de 10 a 30%	de 30% a 50%	acima 50%	TOTAL
SUP	2.233	1.620	3.229	2.407	9.489
MED	1.053	657	1.662	1.274	4.646
FUND	489	284	779	601	2.153
TOTAL	3.775	2.561	5.670	4.282	16.288
P. Consistentes	2.534.571	1.581.399	4.000.434		
	1.177.023	683.588	1.875.053		
	3.400.137	2.121.453			
	1.578.981	917.036			
	1.705.860				
	792.180				
	622.986	361.816	992.446		
	812.718	472.008			
	321.273				
			SOMA	25.950.962	
P Inconsistentes					
	1.467.081	3.711.246	2.844.842		
	634.172	1.739.507	1.342.033		
		2.692.440	2.063.880		
		1.261.980	973.620		
			4.113.746		
			1.940.629		
	299.052	820.287	632.853		
		511.803	394.857		
			998.862		
			SOMA	28.442.890	

A associação entre votação e escolaridade dos candidatos a prefeito em 2016, considerando todos os concorrentes, foi de -0,045. A magnitude do coeficiente é próxima a zero, portanto, nula. Isso indica que não há associação entre o percentual de votos que um candidato obtém e a escolaridade dele. Além disso, o sinal é negativo, indicando uma associação inversa. Quanto maior a escolaridade do candidato, menor a tendência de obtenção de votos dele.

Respostas às Questões do Capítulo III

3.5.1 Saída de resultados

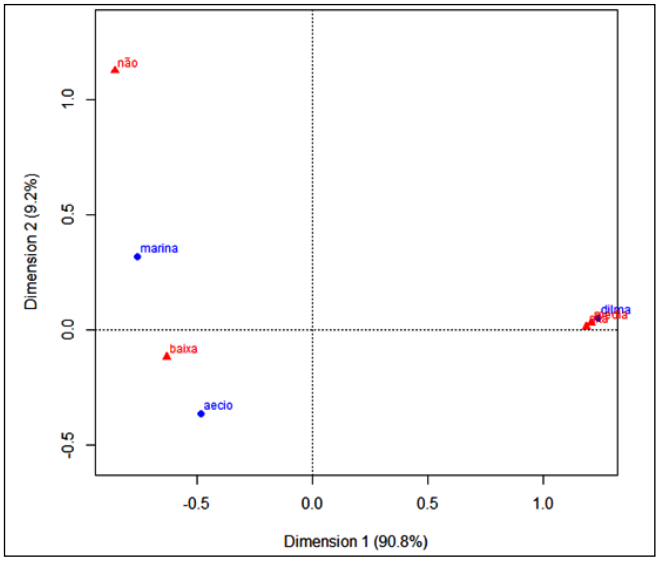
```
reliability(cov(hgpe[,c("AssoAdmDisputa","AssoAdmOutra","postacim-
brig",  "usocargo")], use="complete.obs"))
Rcmdr+    "usocargo")], use="complete.obs"))
Alpha reliability =  0.6514
Standardized alpha =  0.6206

Reliability deleting each item in turn:
                Alpha Std.Alpha r(item, total)
AssoAdmDisputa 0.4331    0.4447         0.6347
AssoAdmOutra   0.7277    0.7499         0.0928
postacimbrig   0.5393    0.4865         0.5382
usocargo       0.5213    0.4441         0.5278
```

O coeficiente fica em 0,651, portanto, acima do mínimo aceitável para a confiabilidade que é de 0,5000. No entanto, os testes considerando a exclusão de uma das variáveis indicam que se for excluída "associação à administração em outra esfera", a confiabilidade sobe para 0,272 e 0,749 no coeficiente padronizado. Sendo assim, a recomendação é montar o índice considerando apenas as outras três variáveis.

3.5.2 Resultados da análise de correspondência canônica:

```
> hgpe = read.table ("clipboard")
       não baixa media alta
dilma    0     1     9    8
marina   3    15     0    0
aecio    0    16     1    1
> ca(hgpe)

 Principal inertias (eigenvalues):
           1        2
Value      0.781291 0.078971
Percentage 90.82%   9.18%

 Rows:
           dilma    marina     aecio
Mass    0.333333  0.333333  0.333333
ChiDist 1.241126  0.819680  0.607057
Inertia 0.513465  0.223958  0.122840
Dim. 1  1.402999 -0.855432 -0.547567
Dim. 2  0.177746  1.126160 -1.303906
```

```
 Columns:
              não     baixa    media     alta
Mass     0.055556  0.592593 0.185185 0.166667
ChiDist  1.414214  0.641957 1.208305 1.186342
Inertia  0.111111  0.244213 0.270370 0.234568
Dim. 1  -0.967786 -0.713790 1.366595 1.342075
Dim. 2   4.007428 -0.421720 0.105264 0.046681
> plot(ca(hgpe))
```

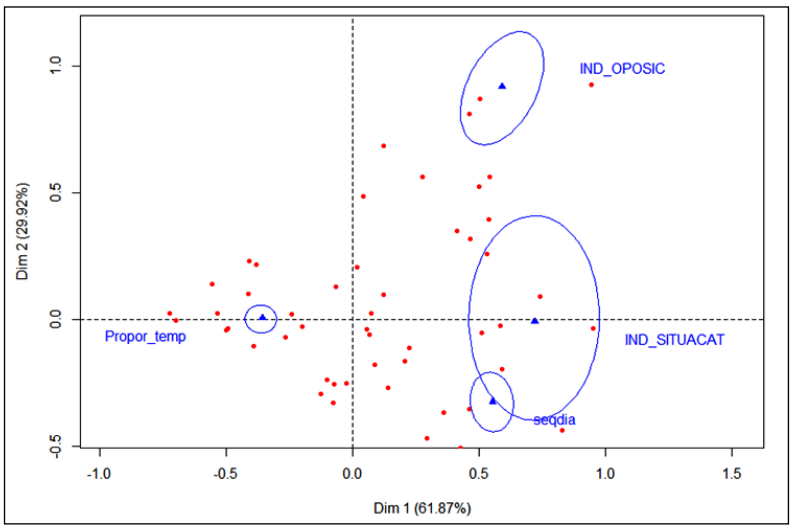


	não	baixa	media	alta	Massa
Dilma	0,184	0,197	0,061	0,055	0,333
Marina	0,184	0,197	0,061	0,055	0,333
Aecio	0,184	0,197	0,061	0,055	0,333
Massa	0,555	0,592	0,185	0,166	

3.5.3 Resultados da análise de correspondência multivariada

```
Call:"res<-CA(hgpe.CA, ncp=5, row.sup=NULL, col.sup=NULL, graph = FALSE)"

The chi square of independence between the two variables is equal to
676.2371 (p-value =  6.324032e-65 ).

Eigenvalues
                       Dim.1   Dim.2   Dim.3
Variance               0.208   0.101   0.028
% of var.             61.872  29.915   8.212
Cumulative % of var.  61.872  91.788 100.000
Columns
             Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr    cos2
seqdia      |   107.930 |   0.553  37.552   0.724 |  -0.326  26.889   0.251 |
Propor_temp |    79.112 |  -0.358  38.004   0.999 |   0.005   0.016   0.000 |
IND_OPOSIC  |   104.947 |   0.590  14.647   0.290 |   0.917  73.092   0.700 |
IND_SITUACAT |    44.199 |   0.720   9.797   0.461 |  -0.008   0.003   0.000 |
```

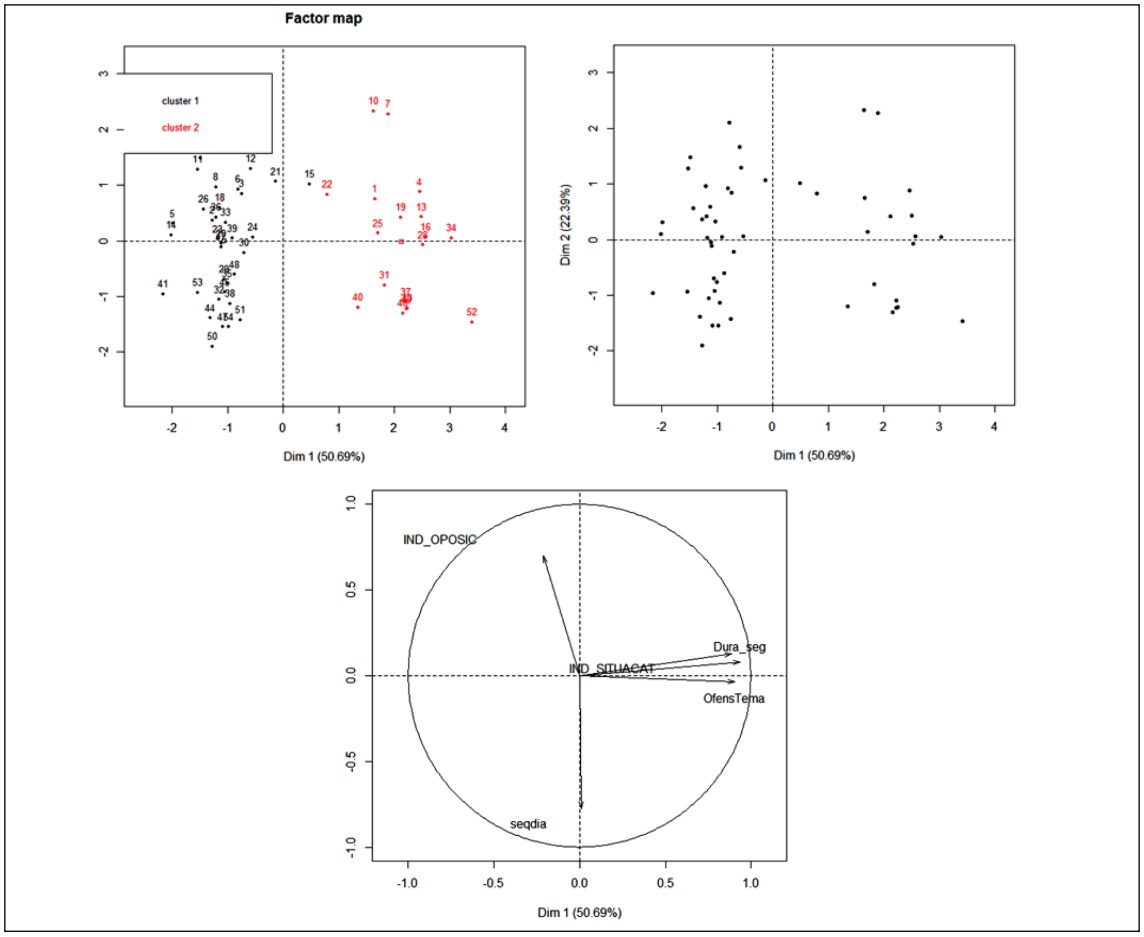
```
                Dim.3     ctr    cos2
seqdia         -0.104  10.055   0.026 |
Propor_temp     0.008   0.160   0.001 |
IND_OPOSIC     -0.105   3.511   0.009 |
IND_SITUACAT    0.779  86.273   0.539 |
RcmdrMsg: [37] NOTA: Os dados hgpe tem 54 linhas e 14 colunas.
```



A correspondência múltipla entre os índices de governismo, oposicionismo, sequência no tempo e duração do segmento mostra proximidade espacial entre índice de situacionismo e dia de exibição, o que indica que o governismo foi crescendo conforme a campanha foi se aproximando do final. O índice de oposição está distante no espaço e a proporção de tempo para os segmentos no dia, também.

3.5.4 Saída dos resultados da análise de componente principais para variáveis do HGPE

```
Call:"res<-PCA(hgpe.PCA , scale.unit=TRUE, ncp=5, graph = FALSE)"

Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
Variance               2.534   1.120   0.868   0.300   0.178
% of var.             50.686  22.391  17.367   6.001   3.554
Cumulative % of var.  50.686  73.078  90.445  96.446 100.000
Variables
               Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
seqdia       |  0.009  0.003  0.000 | -0.777 53.874  0.603 |  0.629 45.556  0.396
Dura_seg     |  0.937 34.656  0.878 |  0.079  0.560  0.006 |  0.001  0.000  0.000
OfensTema    |  0.906 32.385  0.821 | -0.035  0.110  0.001 |  0.038  0.168  0.001
IND_OPOSIC   | -0.212  1.772  0.045 |  0.702 44.010  0.493 |  0.677 52.750  0.458
IND_SITUACAT |  0.889 31.184  0.790 |  0.127  1.446  0.016 |  0.115  1.526  0.013
RcmdrMsg: [50] NOTA: Os dados hgpe tem 54 linhas e 14 colunas.

Rcmdr>  remove(hgpe.PCA)
```

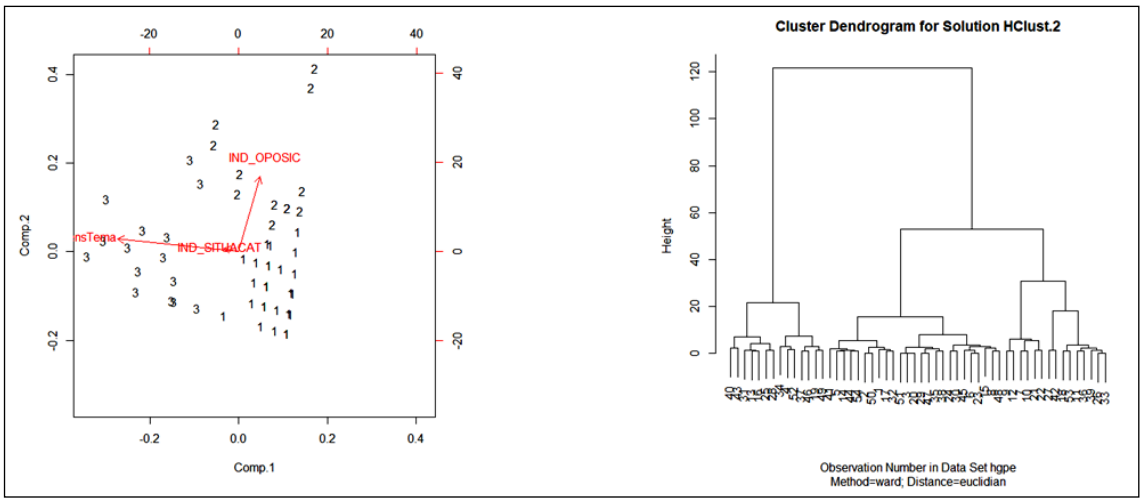


Os componentes principais mostram vetores próximos e na mesma direção para ofensiva quanto a temas, duração média do segmento no dia e índice de situacionismo, indicando que as categorias dessas variáveis tendem a variar na mesma direção. Ofensiva quanto a temas aparece quando há mais governismo e em segmentos com maior duração média. A sequência de dias apresenta outra direção, indicando que não está associada às demais variáveis, assim como o índice de oposicionismo.

3.5.5 Saída de resultados da análise de *clusters* para as variáveis índice de oposicionismo, índice de situacionismo e ofensiva quanto a temas.

```
Rcmdr>  .cluster <-  Kmeans(model.matrix(~-1 + IND_OPOSIC + IND_SITUACAT
+ OfensTema, hgpe), centers = 3, iter.max = 10, num.seeds = 10)

Rcmdr>  .cluster$size # Cluster Sizes
[1] 27 12 15

Rcmdr>  .cluster$centers # Cluster Centroids
  new.x.IND_OPOSIC new.x.IND_SITUACAT new.x.OfensTema
1         1.962963           1.037037        1.518519
2         7.583333           1.416667        2.500000
3         2.133333           2.266667       11.066667
```

```
Rcmdr>  .cluster$withinss # Within Cluster Sum of Squares
[1] 110.6667 160.8333 179.6000

Rcmdr>  .cluster$tot.withinss # Total Within Sum of Squares
[1] 451.1

Rcmdr>  .cluster$betweenss # Between Cluster Sum of Squares
[1] 1237.53

Rcmdr>  biplot(princomp(model.matrix(~-1 + IND_OPOSIC + IND_SITUACAT +
OfensTema,
Rcmdr+    hgpe)), xlabs = as.character(.cluster$cluster))

Rcmdr>  hgpe$Kmeans <- assignCluster(model.matrix(~-1 + IND_OPOSIC +
IND_SITUACAT

Rcmdr>  HClust.2 <- hclust(dist(model.matrix(~-1 +
Rcmdr+    IND_OPOSIC+IND_SITUACAT+OfensTema, hgpe)) , method= "ward")
RcmdrMsg: [57] NOTA: The "ward" method has been renamed to "ward.D";
note new
RcmdrMsg+ "ward.D2"

Rcmdr>  plot(HClust.2, main= "Cluster Dendrogram for Solution HClust.2",
xlab=
Rcmdr+    "Observation Number in Data Set hgpe",
Rcmdr+    sub="Method=ward; Distance=euclidian")
```



O primeiro *cluster* é o maior, reunindo 27 casos. No biplot fica claro que o índice de situacionismo fica entre as outras duas variáveis, com o oposicionismo ficando mais distante de ofensiva em relação a temas. O gráfico dendrograma pelo método Ward indica três grandes *clusters* para os dias de programa eleitoral.

Respostas às Questões do Capítulo IV

4.8.1 Quadro preenchido

	Classes produzidas pelo algoritmo de Reinert					
	CLASSE 1	**CLASSE 2**	**CLASSE 3**	**CLASSE 4**	**CLASSE 5**	**CLASSE 6**
Termos com χ^2 significativo por classe	Experiência, euapoio, unidosporvoce, juntospelobrasil, experiência	Agrotóxico, produto, veneno, ambiente, pl	Predeterminado, destino, candidato, presidência, plano, governo, junto	Silva, lava jato, fake News, tribunal, já	Rede, sustentabilidade, estadual, federal	Aniversário, mulher, luta
% de Ocorrências	17%	18,7%	17%	12,2%	17%	17,9%
Nomes das categorias	Perfil pessoal	Agrícola	Missão	Lava jato	Sustentabilidade	Mulher
% de Postagens	20,5%	6,5%	43,9%	14%	27,1%	14%
% de casos	16,3%	5,2%	34,8%	11,1%	21,5%	11,1%

4.8.2 Planilha preenchida

PÁG.	TIPO	post_message	CL1	CL2	CL3	CL4	CL5	CL6	SOMA
PATRI	link	SAIU NA MÍDIA - UOL NOTÍCIAS O deputado federal Cabo Daciolo (RJ) afirmou que o quadro identificado pelo Atlas da Violência é consequência de "nunca termos enfatizado e valorizado educação no país". "Quanto menor for o seu nível de conhecimento mais fácil fica de escravizar um povo" comentou o parlamentar. #daciolo #patriota #eleicoes2018	0	0	0	0	1	0	1
PATRI	link	SAIU NA MÍDIA - BNC BRASIL NORTE COMUNICAÇÃO #japonesdafederal #patriota #amazonas #manaus	0	0	0	0	1	0	1
PATRI	video	O Patriota faz 6 anos neste 19 de junho é já e maior que quase todos partidos velhos grandes que estão aí hoje quando tinham 6 anos. Estamos no caminho de fazer um partido grande para ajudar nosso Brasil e os brasileiros. Feliz aniversário Patriotas.	0	0	0	0	0	1	1
PATRI	link	No último dia 16 de junho às 9h30 foi realizado a Pré-Convenção do Patriota 51 em São Paulo que contou com a presença de pré-candidatos a Deputado Estadual e Federal Senador membros da executiva estadual e nacional autoridades e outros convidados. #patriota #eleições2018 #sãopaulo	0	0	1	0	1	0	2
PATRI	link	No sábado do dia 16 às 19h00 na Assembleia de Deus – Ministério Alpha com Igreja lotada em Registro-SP o Apóstolo Paulo Corrêa juntamente com Ministros Presbíteros e demais obreiros uniram forças para interceder em oração ao Mestre Jesus pelos membros irmãos Adilson Barroso – presidente nacional do Patriota 51 - que irá disputar uma vaga na Câmara Federal; e Pastor Paulo Corrêa Junior Deputado Estadual que irá concorrer à reeleição. #adilsonbarroso #paulocorrea #patriota #assembleiadedeus	0	0	0	0	1	0	1
PATRI	photo	Hoje é aniversário do nosso querido Presidente Adilson Barroso que Deus possa te dar muita saúde e força nessa batalha em conquista de um Brasil melhor pra todos nós. Parabéns Adilson! #AniversáriodoPresidente #Adilson #patriota51	0	0	0	0	0	1	1
PATRI	photo	Estamos no caminho certo! #Aniversário #Patriota51 #19deJunho	0	0	0	0	0	1	1

PATRI	link	Em defesa da mulher Cássia Barroso falou sobre o espaço da mulher nos lares no trabalho e principalmente na política. Citou as conquistas criticou a discriminação falou sobre projeções futuras elogiou o parecer do TSE em relação a deliberação dos 30% do fundo partidário para uso em campanha e aumento do tempo de TV para as mulheres. #cassiafreire #mulheresnapolitica #patriotamulher #saopaulo	0	0	0	0	0	1	1
PATRI	link	Candidato do Patriota obteve 8.152 votos ou 57 93% dos votos válidos. Na cidade houve 19 68% de abstenção 3 41% de brancos e 9 75% de nulos. SAIU NA MÍDIA - Portal Globo /G1	0	0	1	0	0	0	1
PATRI	photo	#Aniversario #Patriota51	0	0	0	0	0	0	0
PATRI	link	Já está no senado minha proposta de lei que reforça a proibição de descarte de lixo em lugares públicos. Hoje é o Dia Mundial do Meio Ambiente que começou a ser comemorado em 1972 com o objetivo de promover atividades de proteção e preservação do meio ambiente e alertar o público e governos de cada país sobre os perigos de negligenciarmos a tarefa de cuidar do mundo em que vivemos . Disse Líliam Sá em sua rede social. #liliamsa #patriota #meioambiente #preservação	0	1	1	0	1	0	3
PSDB	video	Sua ideia para o futuro do país pode fazer parte do plano de governo do nosso pré-candidato à Presidência da República Geraldo Alckmin. Te ouvir é parte essencial para construirmos juntos um novo Brasil. Tem uma ideia? Mande pra gente! https://ideias.geraldoalckmin.com.br Está em São Paulo? Participe do lançamento! http://presenca.geraldoalckmin.com.br/	0	0	1	0	0	0	1
PSDB	photo	Você pode participar do lançamento oficial do Programa de Governo Participativo de Geraldo Alckmin na próxima quinta-feira (28) em São Paulo! Confirme sua presença: alckm.in/presenca	0	0	1	0	0	0	1
PSDB	video	Você certamente vai se identificar com esse vídeo.	0	0	0	0	0	0	0
PSDB	photo	Tudo o que você compartilha é verdade? É bom verificar; Notícias falsas podem interferir na vida das pessoas e até no futuro do país. Por isso o deputado Luiz Carlos Hauly apresentou o projeto de lei 7604/2017 que prevê a aplicação de multa pela divulgação de #FakeNews nas redes sociais. Saiba mais: http://bit.ly/2loxSrE	0	0	0	1	1	0	2
PSDB	video	Todos nós sabemos que o precisa de um choque de gestão e para o deputado Raimundo Matos (CE) nós temos no PSDB um pré-candidato à Presidência da República à altura desse desafio. Veja seu recado! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	photo	Suas ideias vão ajudar a reconstruir o Brasil. Faça parte do lançamento oficial do Programa de Governo Participativo de Geraldo Alckmin amanhã (28) em São Paulo! Confirme sua presença: alckm.in/presenca Mande suas propostas: alckm.in/ideias	0	0	1	0	0	0	1
PSDB	video	Se você acha que mulher merece respeito compartilhe.	0	0	0	0	0	1	1
PSDB	video	Sabatina do Correio Braziliense com o nosso pré-candidato à Presidência da República Geraldo Alckmin está começando agora. Acompanhe! #CBEntrevista #CBnasEleicoes	0	0	1	0	0	0	1
PSDB	video	Quem conhece Geraldo Alckmin há muito tempo como o deputado Marco Tebaldi (SC) afirma: bom senso capacidade de gestão liderança equilíbrio e caráter são suas características principais. Confira sua mensagem! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Quantos políticos você conhece que têm foco em enxugar a máquina pública? Esta foi uma das realizações de Geraldo Alckmin em São Paulo e agora é seu compromisso com o Brasil. Assista à mensagem do deputado Pedro Cunha Lima (PB) sobre os motivos de seu apoio ao nosso pré-candidato à Presidência da República! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Programas sociais de sucesso que atenderam milhões de pessoas em São Paulo é trabalho de Geraldo Alckmin no Estado. Para nós e para o deputado Floriano Pesaro (SP) Geraldo é sinônimo de serviço público de qualidade. Assista! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2

PSDB	video	Por que Geraldo Alckmin? “Quem tem passado pode apresentar o futuro” afirma nosso deputado Ricardo Tripoli (SP). Alckmin já fez muito pelo Brasil e fará ainda mais: com pés no chão diminuirá a máquina pública e resgatará nossa autoestima. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br/	1	0	1	0	0	0	2
PSDB	video	Pensa em votar em branco? O deputado Izalci Lucas (DF) alerta: quem não gosta de política é governado por quem se interessa. Não deixe que outros decidam por você! Precisamos de Geraldo Alckmin pra tirar o Brasil do abandono. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Para sair da crise precisamos de mudanças. Mas não a qualquer custo e de qualquer jeito certo? Preocupado com o futuro o deputado Danilo Forte (CE) destaca que precisamos de alguém com a segurança e o equilíbrio necessários. Para Danilo não há ninguém melhor: Geraldo Alckmin foi vereador prefeito deputado estadual deputado federal governador. Testado e aprovado! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br/	1	0	1	0	1	0	3
PSDB	photo	O trabalhador não pode ficar refém dos sindicatos como nos últimos 70 anos! “Não faz nenhum sentido obrigar o trabalhador ou a empresa a contribuir para quem não os representa” ponderou o deputado Rogério Marinho após o posicionamento do STF em manter opcional o pagamento do imposto sindical. Leia mais: bit.ly/2tCSUr2	0	0	0	0	0	0	0
PSDB	video	O PSDB nasceu da esperança da luta por um Brasil democrático. São 30 anos enfrentando e superando desafios. Contra o atraso somos a modernidade. Contra a falta de gestão somos a eficiência. Contra a desilusão somos o direito de sonhar. Vamos lutar de maneira incansável para termos um país unido a favor do bem comum e livre da corrupção. Conte sempre com o PSDB. Assista ao nosso vídeo comemorativo! #PSDB30anos	0	0	0	0	0	1	1
PSDB	video	O momento exige de todos nós brasileiros grande responsabilidade. A oportunidade de mudar o país está em nossas mãos! Para o deputado Domingos Sávio (MG) é fundamental termos alguém confiável responsável e capaz como Geraldo Alckmin. Veja seu recado! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	O deputado Carlos Sampaio (SP) dá motivos sólidos para apoiar de forma irrestrita Geraldo Alckmin: um homem sério competente e que não à toa foi governador de São Paulo 4 vezes. Nestes tempos difíceis o que mais precisamos é de um líder testado e aprovado por tantas vezes. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br/	1	0	1	0	0	0	2
PSDB	video	O Brasil passa por uma grave crise e nas palavras do deputado Antonio Imbassahy (BA) precisamos de um timoneiro um presidente confiável e equilibrado que possa fazer o país voltar a crescer. Precisamos de Geraldo Alckmin! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Nosso Deputado Samuel Moreira fez um alerta importante sobre certos políticos cheios de blá-blá-blá e sem experiência administrativa que fogem de debates porque sabem que passarão vergonha: “Nós não aceitamos flerte com o autoritarismo”. Assista ao vídeo e compartilhe com os amigos!	1	0	0	0	0	0	1
PSDB	video	Nossa PEC que prevê a diminuição do número de senadores deputados estaduais e federais irá tramitar no Congresso Nacional! Representando a bancada federal do PSDB o líder na Câmara Nilson Leitão conseguiu recolher 172 assinaturas para a proposta. O objetivo é reduzir gastos e disponibilizar mais recursos para a população. Veja sua declaração! #ForçaPSDB	0	0	0	0	1	0	1
PSDB	video	Nos últimos anos o Brasil tem passado por uma crise sem precedente. Mais do que nunca precisamos de um presidente equilibrado competente sério e que saiba administrar. Para o deputado João Gualberto Vasconcelos (BA) é de Geraldo Alckmin que o país precisa! Assista #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Mato Grosso do Sul representando! Dia de festa e união com o governador Reinaldo Azambuja (MS) e Geraldo Alckmin no #MSquedáCerto	0	0	0	0	0	0	0
PSDB	video	Integridade vida modesta interesse público como prioridade vocação para trabalhar com a população... Não faltam motivos para que o deputado Silvio Torres (SP) apoie a pré-candidatura de Geraldo Alckmin #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2

PSDB	photo	Hoje 23 de junho é Dia do Atleta Olímpico Dia do Lavrador e Dia do Migrante: a todos os esportistas lavradores e migrantes nossos parabéns! E a todos os brasileiros aproveitamos a ocasião para sugerir a leitura de um planos mais importantes para cada um de nós: o projeto de Segurança Pública que foi apresentado neste mês por Geraldo Alckmin presidente do PSDB e pré-candidato à Presidência da República. Confira no link https://buff.ly/2Iyt9Up	0	0	1	0	0	0	1
PSDB	photo	Golpe não! PSDB apoia a Lava Jato. Não à CPI da impunidade! Na Câmara Federal 57 deputados do PT assinaram a favor da CPI da Lava Jato para acabar com as investigações. O PSDB só teve uma assinatura - e o parlamentar já pediu para ser excluído da lista. Quem não deve não teme.	0	0	0	1	1	0	2
PSDB	video	Geraldo Alckmin AO VIVO!	0	0	0	0	0	0	0
PSDB	photo	Fugir do debate significa falta de compromisso com a verdade e com cada brasileiro.	0	0	0	0	0	0	0
PSDB	photo	Fomos responsáveis por algumas das mais importantes conquistas sociais políticas e econômicas do Brasil. E queremos que os brasileiros tenham muito mais! bit.ly/2Klb46I #PSDB30anos	0	0	0	0	0	0	0
PSDB	video	Foi lançado o Plano de Governo Participativo do pré-candidato à presidência Geraldo Alckmin e de todos nós! Na segunda-feira começam as reuniões diárias com os grupos temáticos e seus coordenadores. Queremos te ouvir: a sua proposta pode fazer parte do plano de governo e mudar o Brasil! Apresente sua ideia: https://ideias.geraldoalckmin.com.br	0	0	1	0	0	0	1
PSDB	video	FHC é Geraldo. Geraldo é FHC.	0	0	0	0	0	0	0
PSDB	video	Este é um ano de decisões importantes e não dá pra vacilar. O deputado Lobbe Neto (SP) mandou a real em sua mensagem: a gente precisa de alguém com experiência que pensa no social que pensa em você. Assista! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Estamos nessa! Vem com a gente alckm.in/apoio #JuntosPeloBrasil	1	0	1	0	0	0	2
PSDB	photo	Está garantido e por unanimidade! Somos o 1º partido a destinar 30% dos recursos financeiros para candidaturas femininas. "A participação das mulheres eleva a política fortalece a todos nós fortalece a sociedade brasileira" afirmou Geraldo Alckmin presidente nacional do nosso partido. É um momento histórico para todos nós! Leia mais: http://bit.ly/2tJjpKm #MaisMulheresNaPolítica	0	0	0	0	0	1	1
PSDB	video	Entrevista com Geraldo Alckmin AO VIVO no Metrópoles. Acompanhe com a gente	0	0	0	0	0	0	0
PSDB	photo	Encontrou alguma mentira sobre o PSDB ou algum membro do nosso partido? DENUNCIE! Está no ar o nosso canal oficial para receber denúncias de fake news: http://www.psdb.org.br/denuncie/ O nosso caminho é o da verdade!	0	0	0	1	0	0	1
PSDB	video	Em reunião da executiva nacional em Brasília além de celebrar o aniversário de 30 anos do partido aprovamos a destinação de 30% do fundo eleitoral para as mulheres. Por unanimidade! Fomos o primeiro partido a homologar a decisão do TSE sobre o assunto. As mulheres já foram até proibidas de votar. Hoje elas compõem 51 5% da população mas ocupam apenas 10 5% das cadeiras da Câmara. União vontade de lutar pelo país e igualdade. Foi esse o espírito da reunião da executiva do PSDB hoje em Brasília. Um avanço importante em uma luta histórica! #PSDB30anos	0	0	0	0	0	1	1
PSDB	photo	Durante um anúncio sobre a equipe de programa de governo Geraldo Alckmin desafiou Bolsonaro para um debate sobre Segurança Pública. E aí será que ele vai aceitar ou vai correr?	0	0	1	0	0	0	1
PSDB	video	Disse bem o Nilson Pinto (PA): "Em época de crise tudo que não precisamos é de um aventureiro no governo". Precisamos sim da experiência segurança e competência de Geraldo Alckmin: um bom presidente capaz de retomar nossa autoestima e o crescimento do Brasil. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Decidir o futuro do nosso país é algo que precisa ser feito com consciência. Fazendo isso o deputado Fábio Sousa (GO) não precisou pensar muito para dizer sem hesitar: Geraldo Alckmin é o mais preparado. Afinal quem conhece um pré-candidato que entenda tanto de assuntos como economia e segurança pública como ele? #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br/	1	0	1	0	0	0	2

PSDB	video	Como bem lembra o deputado Marcus Pestana (MG) Tom Jobim dizia que o Brasil não é para amadores: chega de radicalismos intolerâncias e bravatas. Após o desastre promovido pelo PT precisamos de serenidade competência experiência. Em outras palavras: precisamos de Geraldo Alckmin. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br/	1	0	1	0	0	0	2
PSDB	video	As mensagens chegaram de todo o país! Um vídeo colaborativo entre tucanos e com a unidade que precisamos para mudar o Brasil. Estamos juntos em um partido com história e preparados para construir o futuro com cada um de vocês. Assista! Compartilhe #PSDB30anos	0	0	1	0	0	0	1
PSDB	photo	As fake news estão aí e precisamos combatê-las! Encontrou um Pinóquio virtual? Use o nosso canal oficial para encaminhar denúncias. Saiba mais http://www.psdb.org.br/denuncie	0	0	0	1	0	0	1
PSDB	video	Após um período difícil para o Brasil alguns buscam por soluções nos extremos mas não podemos passar por outra experiência malsucedida. O deputado Major Rocha (AC) é claro ao dizer: Geraldo Alckmin traz a confiança que o Brasil precisa. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	Alerta o deputado Nilson Leitão (MT): não é momento de buscarmos um salvador da pátria para presidir o país. Precisamos de um gestor com experiência e currículo de ótimos índices na saúde educação segurança pública. Não poderia ser diferente: Nilson apoia Geraldo Alckmin. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br/	1	0	1	0	0	0	2
PSDB	link	A edição do Jornal Nacional de quinta-feira (31) replica uma matéria frágil publicada no jornal O Estado de São Paulo sobre as obras do Rodoanel Norte. Se o jornalismo não foi capaz de explicar adequadamente o assunto cabe a nós tucanos divulgar a mensagem correta! 1 – Não se trata de decisão do Tribunal de Contas da União mas apenas de um relatório de agosto de 2017 escrito por um único auditor e vazado certamente por motivação eleitoral. 2 – Esse relatório continha dois apontamentos flagrantemente errados que já foram até excluídos. Ele nem foi submetido ainda à apreciação dos ministros do TCU. 3 – O jornalismo apressado no entanto faz os leitores e telespectadores acreditarem que o documento traz a bênção de todos os ministros do tribunal. 4 – O trecho Norte do Rodoanel é a única obra rodoviária em andamento no Estado de São Paulo que conta com repasses federais. 5 – Para cercar-se de todos os cuidados e fazer do trecho um modelo de contratação foi feita uma licitação internacional que seguiu as exigências do Banco Interamericano de Desenvolvimento. 6 – A Dersa no que se refere à sua alçada seguiu a lei e as regras da licitação internacional e já prestou todos os esclarecimentos devidos. 7 – O julgamento demonstrará a regularidade dos procedimentos. COMUNICAÇÃO DO PSDB	0	0	0	1	0	0	1
PSDB	video	A crise se aprofunda há muito tempo. Para sairmos dela é preciso conciliar razão emoção e a vontade de ver o Brasil melhorar. Segundo a deputada Yeda Crusius (RS) isso se faz a partir de governantes éticos e decentes. E por isso quer ver Geraldo Alckmin presidente. #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	87 anos fundador do PSDB estadista presidente do Brasil por dois mandatos o cara da estabilidade econômica e outros grandes feitos ao país além de sua contribuição ao mundo acadêmico como professor e intelectual: hoje é aniversário do grande Fernando Henrique Cardoso. Parabéns FHC	0	0	0	0	0	1	1
PSDB	video	“O maior valor de Geraldo Alckmin é o fato de ser uma pessoa honrada um homem honesto e simples”. Para o deputado Eduardo Cury (SP) Geraldo é o homem certo para comandar o país. Assista sua mensagem! #JuntosPeloBrasil #UnidosPorVocê https://euapoio.geraldoalckmin.com.br	1	0	1	0	0	0	2
PSDB	video	“Ainda falta muito para o brasileiro se sentir seguro ao sair de casa. Para isso defendo que o Governo Federal lidere o combate ao crime organizado com maior proteção das fronteiras e mais participação dos municípios. Queremos uma vida melhor e mais tranquila.” - Geraldo Alckmin presidente nacional do PSDB e pré-candidato à Presidência da República	0	0	1	0	1	0	2
PSDB	video	#JuntosPeloBrasil Hoje São Paulo abraçou Geraldo Alckmin literalmente Vem gente https://euapoio.geraldoalckmin.com.br/	0	0	1	0	0	0	1

REDE	photo	BOLA NA REDE Não dá mais pra tolerar tanta coisa errada no Brasil mas agora é hora de torcer para nossa seleção trazer o Hexa. Vem ai o Bola na REDE em que mobilizadores apoiadores lideranças e pré-candidatos se reunirão para acompanhar os jogos da seleção na Copa e torcer pelo Brasil com Marina. Faça parte dessa festa! Veja os encontros já marcados: http://bit.ly/BlnRede Seja um organizador: http://bit.ly/BlnRede Os encontros serão atualizados a medida que forem sendo confirmados. Vista a camisa! É possível se divertir torcer e fazer a boa política juntos!	0	0	1	0	1	0	2
REDE	video	Pressione os deputados e ajude a impedir a flexibilização no controle de agrotóxicos que fazem mal à saúde. Assine a petição: https://tinyurl.com/l8l5xjl	0	1	0	0	0	0	1
REDE	photo	Política e criatividade...	0	0	0	0	0	0	0
REDE	photo	Parabéns ao deputado Miro Teixeira pelo aniversário que nesse novo ano de vida se renove também sua valorosa contribuição para o bem público as boas leis e a justiça em nosso país.	0	0	0	0	0	1	1
REDE	photo	Os gastos com fretes aéreos contratados por partidos políticos custaram mais de R$ 4 5 milhões em 2017. Você concorda com isso?	0	0	0	0	0	0	0
REDE	photo	Os cabo-frienses começaram a fazer a Operação Lava Voto e sinalizaram um movimento de mudança que pode se espalhar por todo o país em outubro. Parabéns Dr. Adriano pela vitória na eleição suplementar na prefeitura Cabo Frio neste domingo!	0	0	0	1	0	0	1
REDE	photo	O último domingo (24) foi um dia histórico para todos nós! Depois de décadas de luta e muita repressão as mulheres na Arábia Saudita passaram finalmente a ter o direito dirigir seus próprios veículos. O país era a única nação em todo o mundo que ainda mantinha essa restrição às cidadãs. Nosso reconhecimento às mulheres sauditas que com determinação e coragem atuam para construir um país com menos desigualdades entre gêneros.	0	0	0	0	0	1	1
REDE	link	O TSE decidiu em caráter liminar a retirada de cinco postagens mentirosas que associam Marina Silva com as investigações da Lava Jato. A decisão proferida pelo ministro relator Sérgio Banhos confirma o que todos os documentos da Justiça já comprovaram: que a pré-candidata da REDE à presidência não é investigada pela Operação Lava Jato e jamais recebeu propina ou caixa 2 em suas campanhas.	0	0	1	1	1	0	3
REDE	photo	O melhor senador do Brasil lançou sua pré-candidatura à reeleição. Estamos juntos Randolfe Rodrigues!	0	0	1	0	0	0	1
REDE	link	O lançamento da pré-candidatura do professor de direito ambiental da UFSC Rogério Portanova ao governo de Santa Catarina ocorrerá nesta quinta-feira dia 21 às 11h na aldeia guarani Yynn Moroti Wherá em Biguaçu. Na ocasião também serão anunciadas as pré-candidaturas da ambientalista Miriam Prochnow e do consultor em relações internacionais Diego Mezzogiorno ao Senado Federal. Saiba mais em https://goo.gl/Y5ZGGJ	0	0	1	0	1	0	2
REDE	video	O impacto de décadas de desmatamento na Amazônia causam recordes históricos de queimadas e incêndios florestais. Nesta semana do meio ambiente entenda as razões que tornaram a floresta mais inflamável e o que podemos fazer para evitar consequências que comprometem a saúde da populaçao do meio ambiente e a economia. #SemanaDoMeioAmbiente	0	1	0	0	0	0	1
REDE	video	O governo Temer mais uma vez usa as Forças Armadas para esconder as suas incompetências.	0	0	1	0	0	0	1
REDE	photo	O enjaulamento e separação das crianças imigrantes de seus pais presos nos EUA é cruel e indefensável. 49 crianças brasileiras estão isoladas com mais 2000 crianças de todo o mundo. Tirar delas esse vínculo é a perversão da liberdade e dos direitos individuais tão valorizados nos EUA.	0	0	0	0	0	0	0
REDE	photo	O deputado foi o primeiro condenado pela Operação Lava Jato no Supremo Tribunal Federal (STF). Ele pegou 13 anos de prisão em regime fechado por corrupção passiva e lavagem de dinheiro. O parlamentar é acusado de ter recebido propina de R$ 29 milhões de esquema da Petrobrás.	0	0	0	1	1	0	2

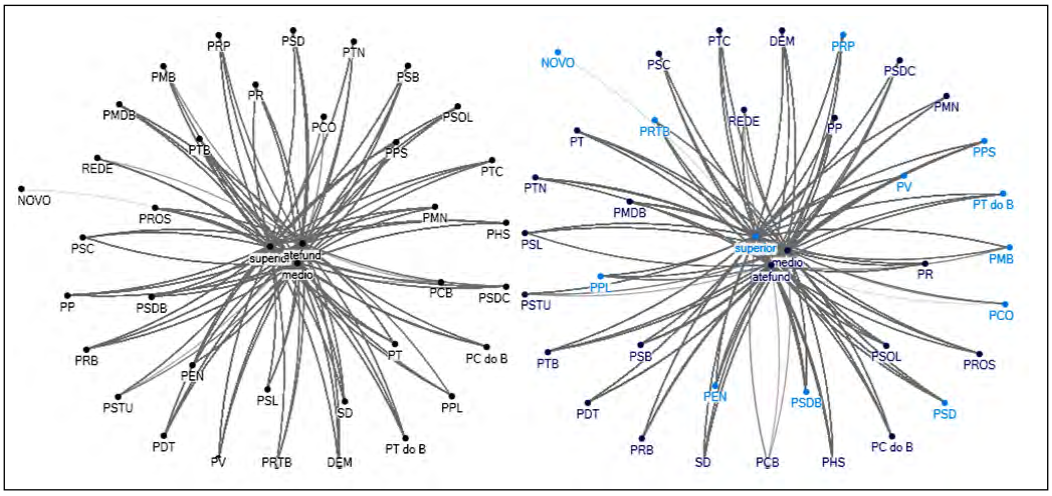
REDE	link	O Brasil tem assistido a uma intensa mobilização dos caminhoneiros em greve devido às seguidas altas no preço dos combustíveis. Além dos caminhoneiros autônomos com demandas legítimas existe a suspeita de um lock-out organizado por empresários da área de transporte para auferirem vantagens durante o processo de negociação. A Rede Sustentabilidade destaca: 1. Que o atual governo cercado de investigações de corrupção demonstra insensibilidade e falta de agilidade em negociar com o segmento dos transportes. Desde outubro representantes dos caminhoneiros vêm tentando diálogo com o governo sem sucesso. Em face do impacto da paralisação no abastecimento dos mais variados produtos e serviços as reações e tomadas de decisões se mostraram lentas frente a processos de mobilização diversos que utiliza modernas tecnologias de comunicação; 2. Que acima da questão entre liberalismo econômico e intervencionismo do Estado no preço da gasolina a crise dos combustíveis está centrada na falta de uma reforma tributária que destrave o setor produtivo sem comprometer a arrecadação do governo; 3. Que uma maior diversificação da matriz energética para além da utilização dos combustíveis fósseis com vistas à diminuição da poluição e da vulnerabilidade externa do país ao preço do petróleo somada à construção de um modal de transportes mais eficiente seja através de ferrovias hidrovias ou outras possíveis formas em nosso país continental; 4. Que a situação atual demonstra a enorme fragilidade do sistema de distribuição de alimentos e da segurança alimentar e hídrica do país. O modelo de produção e consumo valoriza produtos que viajam longamente antes de chegarem à mesa do consumidor e o tratamento e distribuição de água é dependente em muitos lugares dos mesmos combustíveis que fazem o transporte de alimentos. 5. Que os desafios do país são complexos e os problemas interligados exigindo um amplo debate nacional sobre nosso futuro pensando em soluções estruturais e não empurrando os problemas de um setor para outro sem solucioná-los realmente. Temos este ano uma preciosa oportunidade para o exercício democrático de apresentação de propostas para o país e em especial para chegar a um desenvolvimento justo e sustentável. Saudações de luta e de paz!!!	0	1	1	0	1	1	4
REDE	link	No encontro os porta-vozes nacionais da REDE Sustentabilidade Pedro Ivo e Laís Garcia conversaram com o presidente da Fundação da Ordem Social do PROS Felipe Espírito Santo sobre os principais eixos do programa eleitoral e programático das duas legendas. Saiba mais em https://goo.gl/aZYsEP	0	0	0	0	1	0	1
REDE	photo	Neste domingo o Tocantins pode começar a mudar a sua história. Juntos podemos levar o criador da Lei da Ficha Limpa para o 2º turno. Márlon Reis.	0	0	1	0	0	0	1
REDE	photo	Neste Dia Mundial do Meio Ambiente nós da REDE reafirmamos nosso compromisso político com o ideário da sustentabilidade e nossa convicção de que esta agenda precisa ser tratada como prioridade. Economia e ecologia devem ser aliadas - https://bit.ly/2JzD5ut	0	1	0	0	1	0	2
REDE	photo	Nesta sexta (01) a Fundação Rede Brasil Sustentável debaterá Saúde e Direitos Humanos. Compareça ou acompanhe a transmissão pela internet. #18Eixos	0	0	0	0	1	0	1
REDE	photo	Nesta quinta (24) Marina será sabatinada pelo UOL Folha e SBT com transmissão ao vivo no facebook. Acompanhem!	0	0	0	0	0	0	0
REDE	photo	Marina vai se firmando como a melhor opção para unir o Brasil. Nosso projeto é lutar pela construção de um país democrático sustentável que enfrente as desigualdades dê oportunidades para todos e derrote a velha política.	0	0	0	0	0	1	1
REDE	photo	Marina Silva lidera a intenção de voto na Bahia. Fonte: P&A - Pesquisa e Análise	0	0	0	1	0	0	1
REDE	video	Marina Silva é sabatinada pelo UOL Folha e SBT. Compartilhe!	0	0	0	1	0	0	1
REDE	video	Marina ao vivo na Jovem Pan de Curitiba.	0	0	0	0	0	0	0
REDE	video	Hoje nós estamos torcendo pelo mesmo time. Com a palavra o presidente do Flamengo Eduardo Bandeira de Mello.	0	0	0	0	0	0	0
REDE	photo	Hoje é dia de desejarmos felicidades para o grande João Derly líder da Bancada da REDE na Camara Federal. João que o ano que se inicia seja pleno de realizações e de harmonia para você e sua família. Estamos juntos com você em frente nas nossas jornadas de Luta e de Paz!	0	0	1	0	1	1	3
REDE	photo	Hoje à noite Marina estará no Band Eleições a partir das 00h25. Assista e avise seus amigos!	0	0	0	0	0	0	0

REDE	photo	Hoje dia 17 de Maio completam 28 anos que a Organização Mundial de Saúde (OMS) retirou a homo e a bissexualidade do Código Internacional de Doenças passando a reconhecer que não existe cura para o que não é doença. Desde então assistimos a muitos avanços em direitos e na aceitação da diversidade em muitos países mas ainda há muito o que ser conquistado para a plena cidadania e emancipação das pessoas LGBTs. Comportamentos homossexuais ainda são considerados crime em 72 países e a punição em 13 deles é a pena de morte. A OMS ainda considera a transsexualidade como transtorno mental embora sua retirada do CID esteja prevista para este ano. O Brasil é o país que mais mata pessoas LGBTs sendo uma morte a cada 19 horas! O combate a todo tipo de discriminação em decorrência de cor classe religião orientação sexual e identidade de gênero é um compromisso da REDE Sustentabilidade temos em nossos quadros pessoas comprometidas com a liberdade e igualdade a todos. Em tempos onde o ódio vibra nos debates públicos que a nossa palavra de ordem seja "amor". Amar é um direito humano!	0	0	0	0	1	0	1
REDE	photo	Heloisa Helena a Rede Sustentabilidade deseja um novo ano de vida com missões a altura de sua energia compromisso e amorosidade.	0	0	0	0	1	0	1
REDE	video	Essa infâmia de que o marido de Marina Silva é desmatador na Amazônia é uma fake news que já foi desmentida pela Procuradoria Geral da República em 2013. Compartilhe a verdade!	0	0	0	1	0	0	1
REDE	video	Envie suas perguntas e compartilhe!	0	0	0	0	0	0	0
REDE	photo	Compartilhe a lista dos 18 deputados que votaram a favor do PL do Veneno que flexibiliza o uso de agrotóxico.	0	1	0	0	0	0	1
REDE	photo	Celebramos os 56.952 votos que Márlon Reis conquistou na eleição suplementar para o governo do Tocantins. Com poucos recursos e sem grandes estruturas o criador da Ficha Limpa inaugurou um grande movimento. Estamos juntos!	0	0	1	0	0	0	1
REDE	photo	Acompanhe Marina Silva ao vivo na Sabatina dos pré-candidatos realizada pelo UNICURITIBA - Centro Universitário Curitiba: http://bit.ly/marinanaUnicuritiba	0	0	1	1	0	0	2
REDE	link	A REDE vem por meio de seus porta-vozes nacionais lamentar a perda da nossa valorosa Jaqueline Vendruscolo fundadora da REDE em Guaíra-PR e apresentar à família e aos parceiros de jornada política na construção de nosso sonho de país a solidariedade da direção e de todos os nossos filiados nesse momento de perda. Que a saudade nos mantenha a memória da luta pessoal e coletiva de Jaqueline.	0	0	0	0	1	1	2
REDE	link	A Rede Sustentabilidade solidariza-se com o povo da Venezuela no momento político atual em que a democracia está sendo cotidianamente aviltada por um processo político controlado pela força das armas do exército e das milícias. A resposta foi a alta abstenção no processo político-eleitoral (54% das 20 5 milhões de pessoas registradas para votar) que reelegeu o Presidente Nicolás Maduro o que aponta para um sério comprometimento da democracia venezuelana. A Rede Sustentabilidade afirma a importância de retomada do amplo processo democrático com garantia plena de direitos civis e políticos salientando que é preciso superar a crise institucional existente com reconstrução da estabilidade política econômica e social da nação. Por fim o papel do Governo brasileiro deve ser de envidar esforços diplomáticos para solução pacífica da crise interna venezuelana. Executiva Nacional da Rede Sustentabilidade	0	0	1	0	1	0	2
REDE	link	A Rede Sustentabilidade em Santa Catarina fez no último sábado (12) na aldeia guarani Yynn Moroti Wherá ("reflexo das águas cristalinas") também conhecida como M'Biguaçu o primeiro encontro estadual de seus pré-candidatos. Na ocasião o partido confirmou a pré-candidatura de Rogério Portanova a governador do estado. Saiba mais:	0	0	1	0	1	0	2
REDE	photo	A Rede Sustentabilidade de Santa Catarina lançou oficialmente ontem (21) as pré-candidaturas do professor Rogério Portanova ao governo do Estado e da ambientalista Miriam Prochnow e do consultor em relações internacionais Diego Mezzogiorno ao Senado. Os anúncios foram feitos na aldeia guarani Yynn Moroti Wherá em Biguaçu. Saiba mais em https://goo.gl/cjQKpn	0	0	1	0	1	0	2

REDE	link	A REDE SUSTENTABILIDADE conclama o Senado Federal a rejeitar a aprovação do PL 6299/2000 e repudia as tentativas de atender a interesses tão nocivos e com consequências tão danosas para o povo brasileiro. O PL 6299/2000 busca facilitar o registro e o comércio de Agrotóxicos em nosso país. Ele é parte das estratégias das gigantes multinacionais fabricantes desses produtos para ampliar suas posições em nosso mercado. Em contrapartida elas financiaram campanhas para eleger parlamentares. Estão previstos no PL a substituição do nome "agrotóxico" por "produto fitossanitário" bem como a retirada da caveira da embalagem – sinal internacional de perigo que ajuda aos que não sabem ler. São formas de esconder o perigo que estes produtos podem representar à saúde das pessoas e ao meio ambiente cuidando-se apenas de aumentar o mercado de consumo. O Brasil é o maior importador de agrotóxicos do planeta: consome 1/4 de todos os venenos e pesticidas produzidos no mundo. Além de exercerem forte lobby para alterar a legislação local conseguindo a liberação pela ANVISA de pelo menos 14 tipos de substâncias proibidas na maioria dos países da Europa e de negarem os impactos dos agrotóxicos na saúde da população e no meio-ambiente as mesmas empresas se apropriam da agrobiodiversidade pela concentração produtiva agroindustrial: hoje 75% dos alimentos do planeta provém de apenas 12 espécies vegetais e apenas 5 espécies animais. Recentemente numa estratégia de aumento de poder empresarial ocorreu a fusão da Bayer com Monsanto o que torna ainda mais aguda a concentração produtiva do mercado agroquímico com quatro empresas dominando 65% do comércio de agrotóxicos e pesticidas e quase 60% das vendas de sementes no mundo. Com a proibição de alguns dos agrotóxicos mais perigosos em países ricos da Europa está havendo uma corrida para a venda dos mesmos na África e Américas "desovando" estoques e rentabilizando fórmulas e fábricas em flagrante desrespeito às populações e ao meio ambiente. A ação dos deputados brasileiros que apoiam esse PL facilitando o registro dos produtos está a serviço desses interesses e não da população brasileira. Conforme a ANVISA o país tem consumido atualmente uma média de 7kg de veneno ao ano por pessoa sob a complacência da nossa legislação atual. Se a flexibilizarmos ainda mais nossos custos com a saúde subirão. Conforme pesquisadores franceses há aumento no número de cânceres o que levou a França a banir de seu território o glifosato fabricado pela Monsanto. No Brasil estudo realizado na USP demonstra que há riscos de que um doador de sangue da área rural contamine os receptores de seu sangue. Assim como poderão ser mais contaminados ainda os solos a chuva as águas e os alimentos pois a disseminação de agrotóxicos é incontrolável. Conclamamos todos a pressionarem seus parlamentares para a NÃO APROVAÇÃO do PL do Veneno e da morte conhecida e consentida. Não vamos sacrificar a população de nosso país para que os poderosos grupos empresariais dos agroquímicos ampliem suas contas bancárias e seu patrimônio às custas das populações pobres do mundo.	0	1	0	0	1	0	2
REDE	photo	A REDE Sustentabilidade pelo seu Elo Estadual na Bahia manifesta-se em luto pela passagem do excelentíssimo Dr. Waldir Pires que despediu-se de nós hoje dia 22. Waldir foi uma liderança democrática e inspiradora e seu legado de certo continuará a influenciar a boa política e a democracia. Seus dois mandatos como deputado federal vereador de Salvador e governador pela Bahia foram cumpridos com louvor sem máculas e com relevantes serviços ao povo baiano. Nos solidarizamos com seus familiares e amigos.	0	0	0	0	1	0	1
REDE	photo	A REDE foi o primeiro partido a assinar o Termo de Compromisso contra as fake news. Um tratado feito entre os partidos políticos e o Tribunal Superior Eleitoral - Saiba mais: http://bit.ly/redecontrafake	0	0	0	1	1	0	2
REDE	photo	A REDE comemora a decisão do TSE. Esse é o primeiro passo para garantir a efetividade da lei sobre a cota de candidaturas femininas. Estamos atentos contra qualquer tentativa de reverter essa conquista.	0	0	0	0	1	0	1
REDE	video	Marina defende o fim do foro privilegiado Prisão em 2ª instância Lei da Ficha Limpa Fim da reeleição	0	0	0	0	0	0	0
REDE	video	#MarinaNaJovemPan - Entrevista de Marina Silva ao Jornal da Manhã.	0	0	0	1	0	0	1
REDE	video	Realmente estou sumida da Lava Jato e das páginas policiais afirma Marina em entrevista para o Jornal da Manhã na Jovem Pan!	0	0	0	1	0	0	1

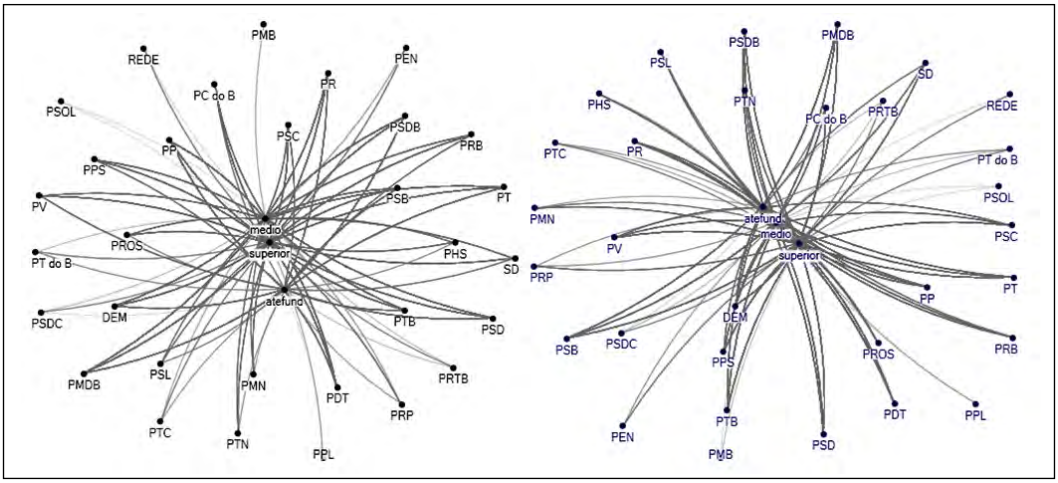
Respostas às Questões do Capítulo V

Grafo para todos os casos, *clusters* por algoritmo de vizinhança Clauset-Neuman-Moore para todos os candidatos. Primeiro é sem *cluster* e o segundo é com os dois *clusters*.



G1	PCB, PSTU, PTC, REDE, SD, PMN, PSC, PP, PSL, DEM, PSB, PDT, PMDB, PTB, PR, PROS, PT, PRB, PTN, PHS, PCdoB, PSDC, PSOL	Até Fundamental Ensino Médio	Dist. Geod. Média = 1,807 Densidade = 0,146
G2	NOVO, PCO, PPL, PRTB, PMB, PEN, PSD, PPS, PTdoB, PRP, PSDB, PV	Ensino Superior	

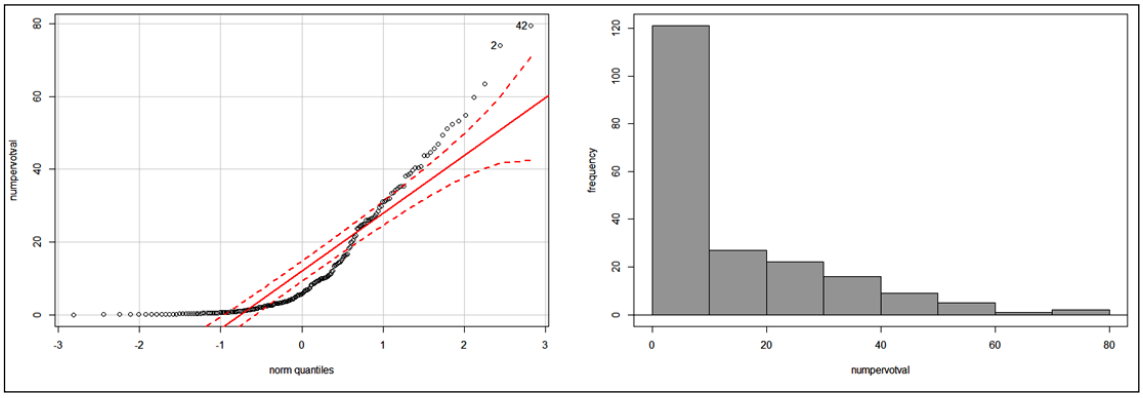
Apenas para os eleitos



G1	PCB, PSTU, PTC, REDE, SD, PMN, PSC, PP, PSL, DEM, PSB, PDT, PMDB, PTB, PR, PROS, PT, PRB, PTN, PHS, PC do B, PSDC, PSOL, NOVO, PCO, PPL, PRTB, PMB, PEN, PSD, PPS, PTdoB, PRP, PSDB, PV.	Até Fundamental Ensino Médio Ensino Superior	Dist. Geod. Média = 1,797 Densidade = 0,156

Respostas às Questões do Capítulo VI

6.7.1 Gráficos de normalidade




```
> kurtosis((Capital$numpervotval))
[1] 4.978976
> Skewness(Capital$numpervotval)
[1] 1.523347
```

Interpretação: Não é possível usar a variável original, pois ela quebra o pressuposto da normalidade de resíduos. A Curtose fica bem acima de 3,000 e o *Skewness* distante de zero. A recomendação é um tipo de transformação para normalizar as distribuições. Pode ser feita pelo método de logaritmo.

6.7.2 Testes de correlação:

Pearson: 0,051, Spearman: 0,040 e Kendall: 0,030.

O teste correto é o de Kendall, pois uma das variáveis é categórica ordinal. As diferenças entre os coeficientes mostram que Kendall é mais restritivo, pois produz o coeficiente mais baixo.

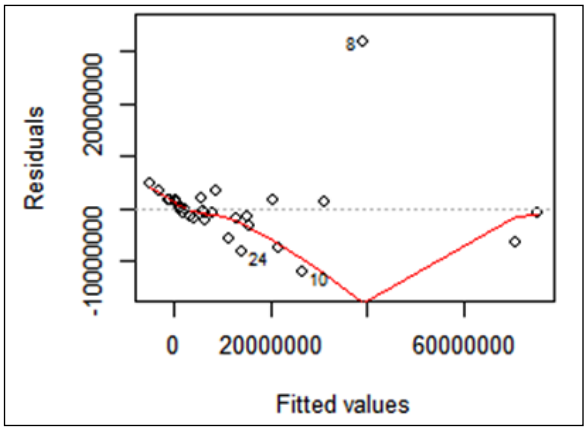
```
Rcmdr>   with(Capital, cor.test(IDADE, VOTCAND_CAT, alternati-
ve="two.sided", method="pearson"))

      Pearson's product-moment correlation

data:  IDADE and VOTCAND_CAT
t = 0.73399, df = 201, p-value = 0.4638
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08662401  0.18807334
```

```
sample estimates:
       cor
0.05170258


Rcmdr>  with(Capital, cor.test(IDADE, VOTCAND_CAT, alternative="t-
wo.sided", method="spearman"))

      Spearman's rank correlation rho

data:  IDADE and VOTCAND_CAT
S = 1338200, p-value = 0.5697
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04013468


Rcmdr>  with(Capital, cor.test(IDADE, VOTCAND_CAT, alternative="t-
wo.sided", method="kendall"))

      Kendall's rank correlation tau

data:  IDADE and VOTCAND_CAT
z = 0.54183, p-value = 0.5879
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
0.03007382
```

Respostas às Questões do Capítulo VII

7.7.1 Os resíduos mostram-se muito distante do esperado, com mais dispersão para valores negativos e a mediana distante de zero. O r² é alto, de 0,884 e a estatística F altamente significativa, em 51,42. Em relação às contribuições individuais, a variável número de doações de pessoas físicas é negativa e todas as demais positivas. No entanto, apenas duas delas tem significância estatística: total de doações em R$ para o partido e total de votos de legenda. O gráfico QQ de resíduos mostra que a distribuição deles não é aleatória, impedindo a extrapolação dos resultados. Como consequência, a estatística *Breusch-Pagan* ultrapassa em muito o limite crítico, não permitindo dizer que há homoscedasticidade na distribuição. O VIF para todas as variáveis fica abaixo de 10,0, o que mostra ausência de colinearidade entre as variáveis explicativa.

```
Rcmdr>   RegModel.18 <- lm(VOTNOM~DOAPFIS+DOAR.TOT+NUMCAND+VOTLEG, da-
ta=legenda)

Rcmdr>  summary(RegModel.18)
Call:lm(formula = VOTNOM ~ DOAPFIS + DOAR.TOT + NUMCAND + VOTLEG,data =
legenda)

Residuals:
      Min         1Q     Median         3Q        Max
-12043918  -1787768   -320882   1632745  31798948

Coefficients:
                Estimate   Std. Error t value   Pr(>|t|)
(Intercept) -5393199.058  3367153.087  -1.602     0.1209
DOAPFIS         -907.337     1037.994  -0.874     0.3898
DOAR.TOT          17.973        7.711   2.331     0.0275 *
NUMCAND        38197.165    23083.547   1.655     0.1096
VOTLEG            17.270        2.842   6.077 0.00000173 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7330000 on 27 degrees of freedom
Multiple R-squared:  0.884,   Adjusted R-squared:  0.8668
F-statistic: 51.42 on 4 and 27 DF,  p-value: 3.05e-12
```




```
Breusch-Pagan test

data:  VOTNOM ~ DOAPFIS + DOAR.TOT + NUMCAND + VOTLEG
BP = 20.93, df = 1, p-value = 0.000004763
Rcmdr>  vif(RegModel.18)
 DOAPFIS DOAR.TOT  NUMCAND   VOTLEG
3.392486 1.557466 2.005184 4.028863

Rcmdr>  round(cov2cor(vcov(RegModel.18)), 3)#Correlations of parameter
estimates
             (Intercept) DOAPFIS DOAR.TOT NUMCAND VOTLEG
(Intercept)        1.000   0.093   -0.454  -0.832  0.366
DOAPFIS            0.093   1.000   -0.153  -0.227 -0.598
DOAR.TOT          -0.454  -0.153    1.000   0.206 -0.335
NUMCAND           -0.832  -0.227    0.206   1.000 -0.377
VOTLEG             0.366  -0.598   -0.335  -0.377  1.000
```

7.7.2 As estatísticas de resíduos do modelo mostram uma mediana próxima a zero e intervalo de valores máximo e mínimo equidistantes de zero, o que indica uma distribuição homoscedástica dos resíduos. As duas variáveis explicativas são positivas e significativas a p<0,050, ou seja, receber doações e ter candidatos contribui para estar no grupo de partidos com mais votos. Porém, os odds-ratio mostram contribuições individuais baixas. No caso de valor de doações: (1,000-1)*100 = 0,002% de chance a mais de ser partido grande para cada Real a mais arrecadado. No caso de número de candidatos: (1-1,047)*100 = 4,7% a mais de chance de estar no grupo dos partidos grandes para cada candidato a mais que o partido possui. Ter mais candidato contribui mais em transformar um partido de pequeno em grande do que arrecadar mais. Isso vale para a eleição de deputado federal em 2014.

```
Rcmdr>  GLM.19 <- glm(PARTGRD ~ DOAR.TOT + NUMCAND, family=binomial(lo-
git),
Rcmdr+    data=legenda)

Rcmdr>  summary(GLM.19)

Call: glm(formula = PARTGRD ~ DOAR.TOT + NUMCAND, family = binomial(lo-
git),
    data = legenda)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.87986  -0.26089  -0.00427   0.16568   1.48812

Coefficients:
                Estimate    Std. Error z value Pr(>|z|)
(Intercept) -10.439936502   4.363162698  -2.393   0.0167 *
DOAR.TOT      0.000020924   0.000008881   2.356   0.0185 *
NUMCAND       0.046177831   0.021224650   2.176   0.0296 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.361  on 31  degrees of freedom
Residual deviance: 11.801  on 29  degrees of freedom
AIC: 17.801

Number of Fisher Scoring iterations: 7

Rcmdr>  exp(coef(GLM.19))  # Exponentiated coefficients (“odds ratios”)
  (Intercept)      DOAR.TOT       NUMCAND
0.00002924106 1.00002092459 1.04726062975

Rcmdr>  vif(GLM.19)
DOAR.TOT  NUMCAND
2.020647 2.020647

Rcmdr>  round(cov2cor(vcov(GLM.19)), 3) # Correlations of parameter es-
timates
            (Intercept) DOAR.TOT NUMCAND
(Intercept)       1.000   -0.863  -0.948
DOAR.TOT         -0.863    1.000   0.711
NUMCAND          -0.948    0.711   1.000
```

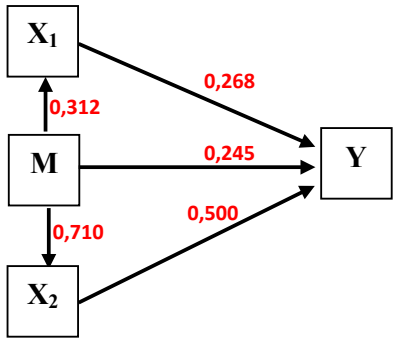
Respostas às Questões do Capítulo VIII

```
lm(formula = TOTVOT14 ~ `DOAR$TOT14` + DOATOT14 + TOTCAND14)
Multiple R-squared:  0.7629,  Adjusted R-squared:  0.732
F-statistic: 24.67 on 3 and 23 DF,  p-value: 2.266e-07

lm.beta(md1)
`DOAR$TOT14`      DOATOT14     TOTCAND14
   0.2687625     0.5002365     0.2456648

lm.beta(md2)DOAR$TOT14
TOTCAND14
0.3125668

lm.beta(md3)DOATOT14
TOTCAND14
0.7103827
```

Efeito direto de doações R$ sobre total de votos: 0,268
Efeito direto de total de doadores sobre total de votos: 0,500
Efeito direto de total de candidatos sobre total de votos: 0,245

Efeito indireto de total de candidatos por total de doações R$: (0,312 x 0,268) = 0,083
Efeito indireto de total de candidatos por total de doadores: (0,710 x 0,500) = 0,355

Var. Moderadora	**Var. Explicativas**	**Var. Dependente**
M. Total de candidatos a deputado federal.	X_1. Total de recursos R$ declarados. X_2. Total de doadores ao partido.	Y. Total de votos ao partido em 2014.

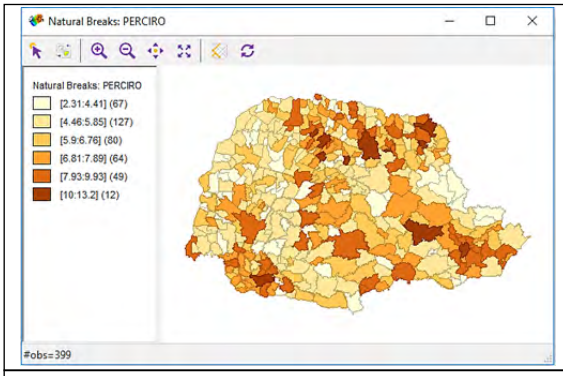


Var. Explicativa	**Efeito Total = (Ef. Direto + (Ef. Indireto))**
X_1. Total de recursos R$ declarados por candidatos em 2014.	$Et_{x1} = a_1 + (b_1 * a_1)$=0,268 + 0,083=**+0,351**
X_2. Total de doadores ao partido em 2014.	$Et_{x3} = a_3 + (b_2 * a_3)$ = 0,500 + 0,355 = **+0,855**
M. Total de candidatos a deputado federal em 2014.	$Et_m = a_2 + ((b_1 * a_1) + (b_2 * a_3))$ = 0,245 + 0,083 + 0,355 = **+0,683**

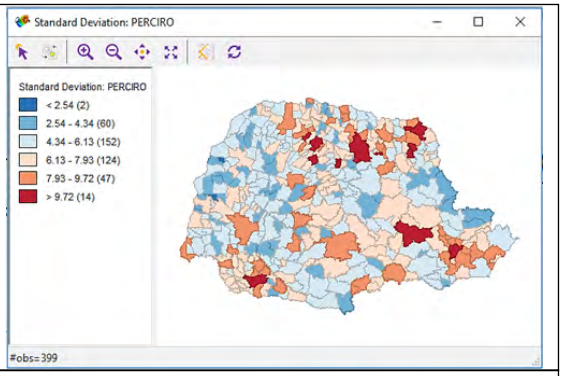
Os resultados mostram que o efeito direto forte é o do total de doadores, seguido de número de candidatos e por último o total de doações em R$. Após a moderação, a diferença entre total de doadores e total de candidatos cai e a diferença entre total de R$ e total de doações aumenta. Isso mostra que o efeito de moderação do número de candidatos é distinto em relação às variáveis explicativas.

Respostas às Questões do Capítulo IX

9.6.1 Gere o mapa por quebras naturais com 6 categorias para a votação de Ciro Gomes (PDT) no Paraná (variável "PERCIRO") e gere o mapa de Desvio Padrão para a mesma variável. Interprete as diferenças entre os dois mapas.



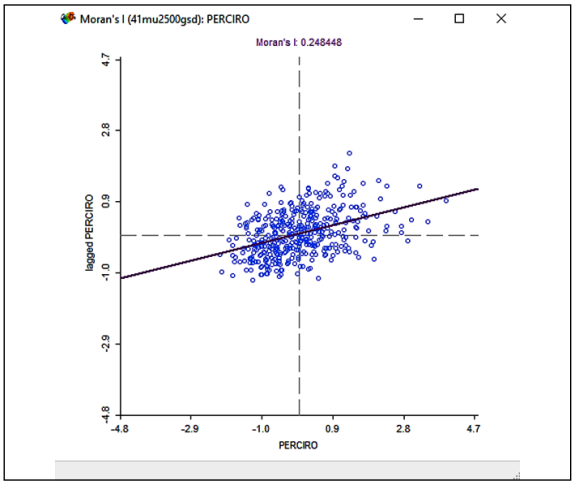
Quebras naturais



Desvio Padrão

INTERPRETAÇÃO: O número de categorias é o mesmo, mas no mapa de quebras naturais há maior número de municípios nas extremidades, enquanto que na distribuição por desvio padrão há uma concentração de municípios nas categorias centrais.

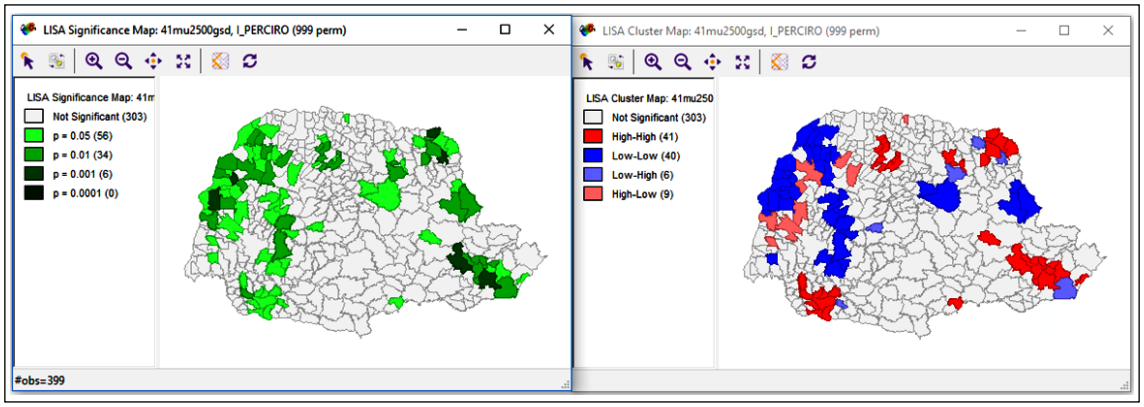
9.6.2 Gere o gráfico de dispersão de I de Moran para a variável "PERCIRO" e interprete o coeficiente.



INTERPRETAÇÃO: O coeficiente de autocorrelação global I de Moran mostra-se baixo, com 0,248, o que significa que não houve um efeito de vizinhança forte para a votação de Ciro Gomes nos municípios do Paraná no primeiro turno das eleições de 2018. Comparando com os coeficientes de Bolsonaro e Haddad, que se encontram no capítulo 9,

Ciro Gomes apresentou um efeito de vizinhança global com metade da intensidade dos dois anteriores.

9.6.3 Gere o gráfico e mapas do coeficiente LISA para a variável “PERCIRO” em municípios do Paraná e interprete os mapas de significância e de *clusters* de LISA.



INTERPRETAÇÃO: o gráfico de significância mostra que em 303 dos 399 municípios não houve significância estatística para *clusters* locais da votação de Ciro Gomes. O gráfico de *clusters* indica que Ciro Gomes apresentou 41 municípios em *clusters* consistentes positivos (alto-alto), que se localizam em diferentes regiões do estado, contra 40 municípios com *clusters* consistentes negativos (baixo-baixo), principalmente nas regiões Oeste e Norte do Paraná.

9.6.4 Identifique qual o melhor modelo de dependência espacial para uma regressão onde a variável dependente é “PERCIRO” e as variáveis explicativas são votação para deputado federal pelo PDT por município do Paraná (F12_PERC) e votação em Ratinho Jr. (G55_PERC). Interprete os resultados.

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : 41mu2500gsd
   (row-standardized weights)
TEST                           MI/DF         VALUE           PROB
Moran's I (error)              0.2102         7.0472         0.00000
Lagrange Multiplier (lag)         1          48.2083         0.00000
Robust LM (lag)                   1           1.9465         0.16296
Lagrange Multiplier (error)       1          46.3956         0.00000
Robust LM (error)                 1           0.1338         0.71452
Lagrange Multiplier (SARMA)       2          48.3421         0.00000
================================ END OF REPORT ================================
```

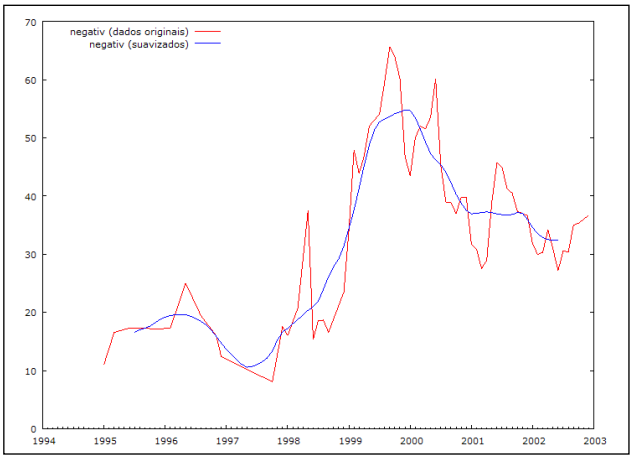
INTERPRETAÇÃO: O diagnóstico de dependência espacial gerado a partir da regressão MQO indicou que, para as variáveis do modelo, a melhor alternativa de dependência é pelo método de defasagem, após comparar os coeficientes de Lagrange de lag (48,20) e error (46,39).

```
REGRESSION
----------
SUMMARY OF Output: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : 41mu2500gsd
Spatial Weight      : 41mu2500gsd
Dependent Variable  :     PERCIRO  Number of Observations:  399
Mean dependent var  :     6.13168  Number of Variables   :    4
S.D. dependent var  :     1.79241  Degrees of Freedom    :  395
Lag coeff.   (Rho)  :    0.390805

R-squared           :    0.196049  Log likelihood        :    -761.618
Sq. Correlation     : -            Akaike info criterion :     1531.24
Sigma-square        :     2.58289  Schwarz criterion     :     1547.19
S.E of regression   :     1.60714

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
         W_PERCIRO      0.390805      0.0635128        6.15317     0.00000
          CONSTANT        5.4782       0.614997        8.90768     0.00000
          F12_PERC     0.0474868      0.0176665        2.68796     0.00719
          G55_PERC     -0.033924     0.00760033       -4.46349     0.00001
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      2         4.6162     0.09945
```

INTERPRETAÇÃO: O coeficiente de ajustamento do modelo é baixo, $r^2 = 0,196$, ou seja, mesmo com dependência espacial, apenas 19,6% das variações de votos em Ciro Gomes são explicadas pelas variáveis. No modelo de erro, o coeficiente espacial (*Rho*) é representado pelo W. É dele o maior efeito individual ($\beta = 0,390$). Depois vem a votação para deputado federal do PDT ($\beta = 0,047$), seguido pela votação de Ratinho Jr. ($\beta = -0,033$). O sinal negativo no coeficiente de Ratinho Jr. indica que as variações foram em sentido inverso. Conforme cresce votação de Ciro Gomes, tende a diminuir a de Ratinho Jr. nos municípios. Além disso, os três coeficientes são estatisticamente significativos (probabilidade < 0,050). Vale ressaltar que o modelo não passou no teste de heteroquedasticidade (prob > 0,050), ou seja, ele quebra o pressuposto da homoquedasticidade e, portanto, não pode ser usado para fazer inferências estatísticas.

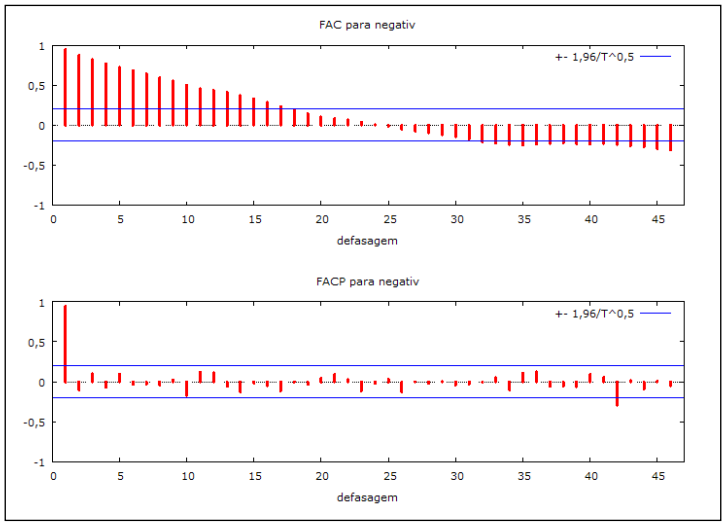
Respostas às Questões do Capítulo X

10.8.a) Gerar o gráfico de médias móveis simples (centradas em 12 meses) para a variável dependente.



RESPOSTA: A série suavizada mostra-se claramente com tendência (suavizada pelas médias móveis) e com sazonalidade (que não desaparece com as médias para 12 meses)

10.8.b) Gerar e analisar os gráficos de autocorreção FAC e FACP, propondo ajustes necessários ao modelo.



RESPOSTA: O correlograma mostra que as FAC apresentam memória, pois a maior parte dos retardos fica acima do limite crítico de 0,050. A FACP mostra que a parcial,

desconsiderando a memória do primeiro retardo, fica praticamente com todos os demais pontos dentro da margem de erro, embora ainda exista uma barra com “valor explodindo” para fora dela.

É possível testar a série com um retorno.

10.8.c) Rodar o teste de Raízes Unitárias ADF para a avaliação negativa de FHC e responder se é possível considerar a série como estacionária ou não.

```
Teste Aumentado de Dickey-Fuller para negativ
testar para baixo a partir de d 11efasagens, critério AIC
tamanho da amostra: 95
hipótese nula de raiz unitária: a = 1

  teste com constante
  incluindo 0 defasagens de (1-L)negativ
  modelo: (1-L)y = b0 + (a-1)*y(-1) + e
  valor estimado de (a - 1): -0,0568198
  estatística de teste: tau_c(1) = -1,75957
  p-valor 0,3983
  coeficiente de 1ª ordem para e: 0,126

  com constante e tendência
  incluindo 0 defasagens de (1-L)negativ
  modelo: (1-L)y = b0 + b1*t + (a-1)*y(-1) + e
  valor estimado de (a - 1): -0,0787078
  estatística de teste: tau_ct(1) = -1,90911
  p-valor 0,642
  coeficiente de 1ª ordem para e: 0,138
```

RESPOSTA: O teste de raiz unitária mostra que nos dois casos, só com constante e com constante e tendência, sem incluir defasagens, o *p-valor* fica muito acima do limite crítico de 0,050. Ou seja, existem raízes unitárias nessa série temporal. Ela não pode ser considerada estacionária.

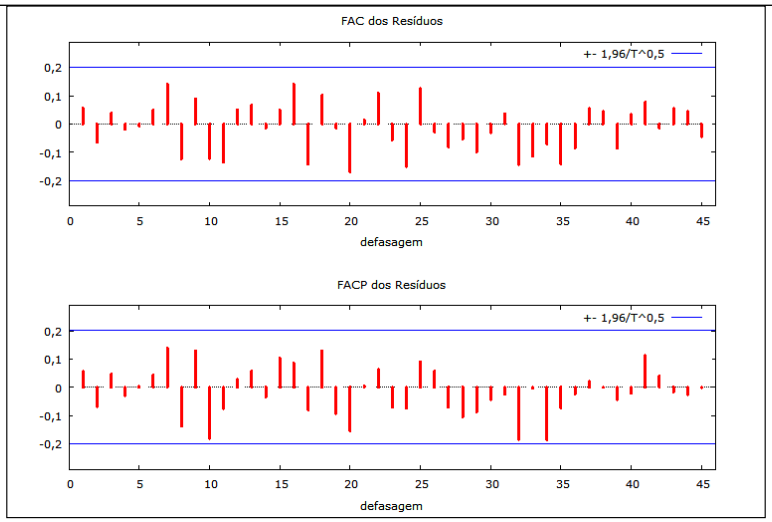
10.8.d) Ajustar o melhor modelo ARIMA (p,d,q), começando por (1,0,1) para a variável dependente e depois ajustando até a retirada do efeito de memória. Rodar um correlograma do modelo para testar a existência de memória.

Modelo: ARMAX, usando as observações 1995:02-2002:12 (T = 95)
Variável dependente: negativ
Erros padrão baseados na hessiana

	Coeficiente	Erro Padrão	z	*p-valor*	
const	27,3782	6,20417	4,413	<0,0001	***
phi_1	0,880317	0,0518362	16,98	<0,0001	***
pri_mand	−14,5368	4,41409	−3,293	0,0010	***
Desemp_1	1,21205	0,600298	2,019	0,0435	**

Média var. dependente	29,55874	D.P. var. dependente	14,92939
Média de inovações	0,069861	D.P. das inovações	4,403663
Log da verossimilhança	−276,3764	Critério de Akaike	562,7528
Critério de Schwarz	575,5222	Critério Hannan-Quinn	567,9126

	Real	Imaginária	Módulo	Frequência
AR				
Raiz 1	1,1360	0,0000	1,1360	0,0000



RESPOSTA: O modelo inicial ARIMA (1,0,1) mostra o coeficiente *theta* não significativo. Então, o modelo deve ser ajustado para (1,0,0), que resulta em todos os coeficientes estatisticamente significativos. O correlograma do modelo (1,0,0) mostra todas as barras de FAC e FACP dentro da margem de erro.

10.8.e) Calcular um modelo autoregressivo (AR1) para descrição das relações ao longo do tempo entre Avaliação negativa de FHC, como dependente, e retorno da taxa de desemprego e se está ou não no Primeiro Mandato, como explicativas. Analisar os resultados.

Modelo 1: Prais-Winsten, usando as observações 1995:02-2002:12 (T = 95)
Variável dependente: negativa
rho = 0,88862

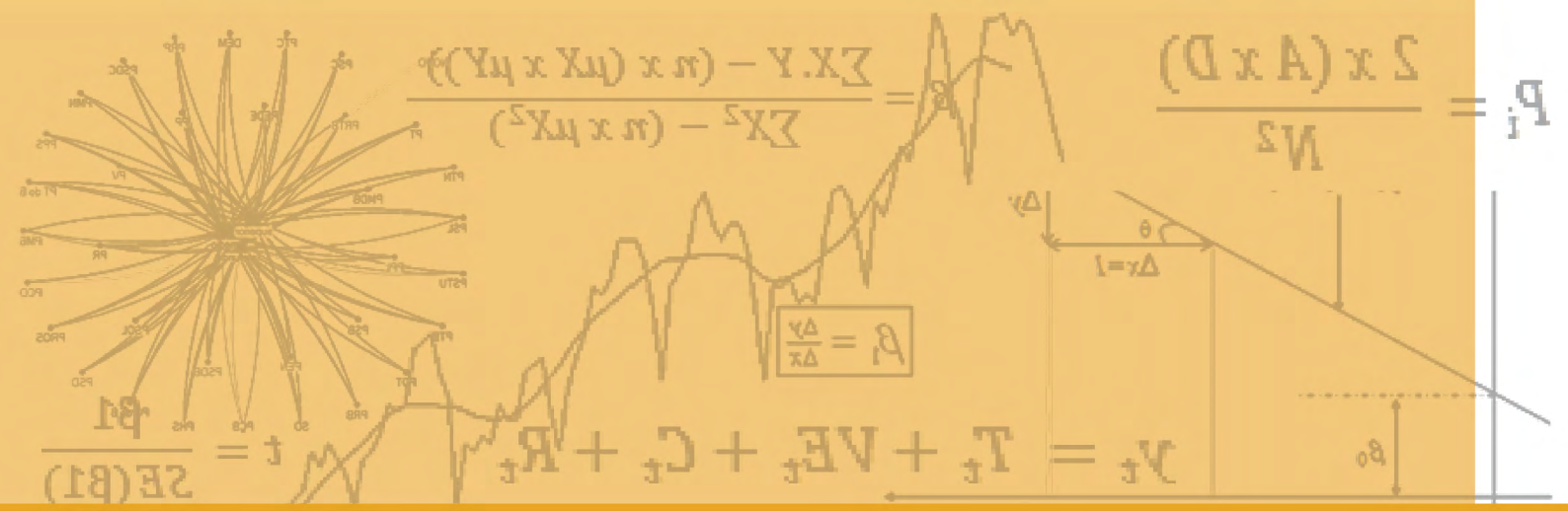
	Coeficiente	Erro Padrão	razão-t	*p-valor*	
Constante	27,1148	6,22504	4,356	<0,0001	***
Primeiro mandato	−14,2079	4,03161	−3,524	0,0007	***
Desemprego_1	1,21928	0,609689	2,000	0,0485	**

Estatísticas baseadas nos dados rô-diferenciados:

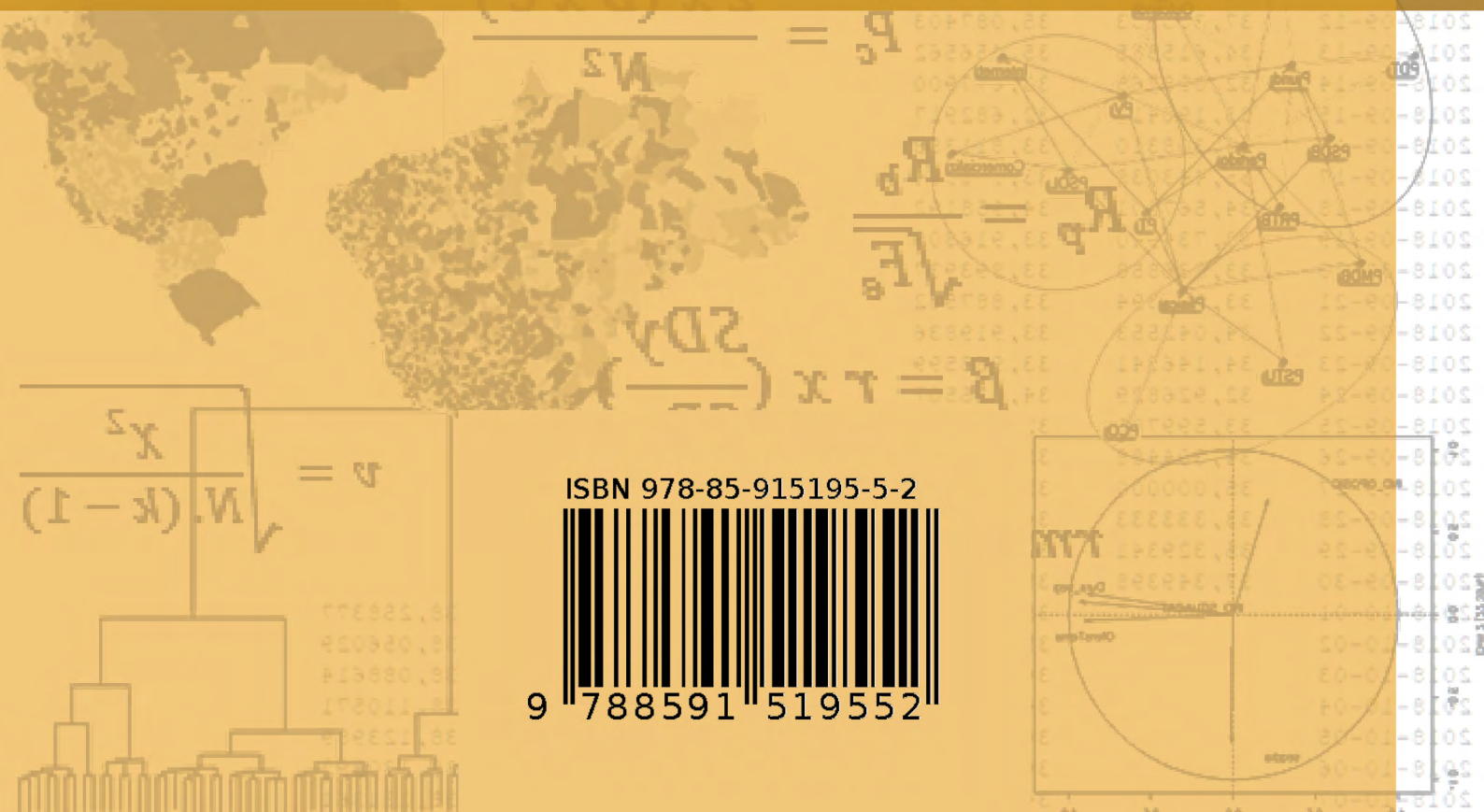
Média var. dependente	29,55874	D.P. var. dependente	14,92939
Soma resíd. quadrados	1841,460	E.P. da regressão	4,473910
R-quadrado	0,912374	R-quadrado ajustado	0,910469
F(2, 92)	7,212837	*P-valor*(F)	0,001231
Rô	0,05178	*Durbin-Watson*	1,893620

RESPOSTA: As estatísticas do modelo mostram um r² de 0,912, indicando bom ajustamento explicativo. E a estatística *Durbin-Watson* fica acima de 1,85, portanto, não indicando existência de raízes unitárias nas variações temporais das séries. Quanto aos impactos individuais, a significância estatística da constante mostra que as variações ao longo do tempo da variável dependente não seguem um padrão único. A variável *dummy* estar ou não no primeiro mandato é a que apresenta uma explicação estatisticamente significativa (*p-valor*=0,000) e coeficiente negativo de -14,20, ou seja, não estar no primeiro mandato aumenta o percentual de avaliação negativa do presidente. Dito de outra forma, a avaliação negativa foi maior no segundo do que no primeiro mandato e as diferenças são estatisticamente significativas. A variável explicativa contínuca “taxa de desemprego com um retardo” também apresentou significância estatística a 95% (*p-valor* = 0.048) e coeficiente positivo de 1,219. Ou seja, conforme aumenta a taxa de desemprego no mês anterior, tende a crescer a avaliação negativa no mês seguinte. Ao longo do capítulo foi mostrado que a avaliação positiva não sofre influência estatisticamente significativa da taxa de desemprego, mas o mesmo não acontece com avaliação negativa que é mais sensível às variações da variável explicativa.

Agência Brasileira do ISBN
ISBN 978-85-915195-5-2



www.cpop.ufpr.br www.cienciapolitica.ufpr.br



ISBN 978-85-915195-5-2