

**UNIVERSIDADE DE SÃO PAULO**  
Instituto de Ciências Matemáticas e de Computação

**Uma Abordagem Estatística para a Análise dos Resultados  
das Eleições Presidenciais**

**Victor Eduardo Lachos Olivares**

Dissertação de Mestrado do Programa Interinstitucional de  
Pós-Graduação em Estatística (PIPGEs)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: \_\_\_\_\_

**Victor Eduardo Lachos Olivares**

## Uma Abordagem Estatística para a Análise dos Resultados das Eleições Presidenciais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Jorge Luis Bazán Guzmán

**USP – São Carlos**  
**Fevereiro de 2024**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi  
e Seção Técnica de Informática, ICMC/USP,  
com os dados inseridos pelo(a) autor(a)

L128a Lachos Olivares, Victor Eduardo  
Uma Abordagem Estatística para a Análise dos  
Resultados das Eleições Presidenciais / Victor  
Eduardo Lachos Olivares; orientador Jorge Luis  
Bazán Guzmán. -- São Carlos, 2024.  
87 p.

Dissertação (Mestrado - Programa  
Interinstitucional de Pós-graduação em Estatística) --  
Instituto de Ciências Matemáticas e de Computação,  
Universidade de São Paulo, 2024.

1. Modelo de regressão Beta. 2. Análise de  
componentes principais. 3. Transformação de razão  
logarítmica. 4. Dados composicionais. I. Bazán  
Guzmán, Jorge Luis , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:  
Gláucia Maria Saia Cristianini - CRB - 8/4938  
Juliana de Souza Moraes - CRB - 8/6176

**Victor Eduardo Lachos Olivares**

**An statistical approach for analysis of presidential elections  
results**

Master dissertation submitted to the Institute of  
Mathematics and Computer Sciences – ICMC-USP  
and to the Department of Statistics – DEs-UFSCar, in  
partial fulfillment of the requirements for the degree of  
the Master Interagency Program Graduate in Statistics.  
*FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Jorge Luis Bazán Guzmán

**USP – São Carlos**  
**February 2024**

*Este trabalho é dedicado à minha família, que me apoiou durante toda a jornada rumo aos meus objetivos. Um agradecimento especial ao meu irmão, que sempre esteve atento e inquebrantável na confiança que depositou em mim; à minha mãe, por ser meu guia e me educar desde a infância; e ao meu pai, que me incentivou a seguir esse caminho.*

# AGRADECIMENTOS

---

---

Gostaria, em primeiro lugar, de expressar minha gratidão a Deus porque através de sua vontade e amor infinito, alcancei este objetivo significativo.

Além disso, quero estender meu agradecimento à minha família: minha mãe, Rosalina Emma; meu irmão, Carlos Alberto; e meu pai, Victor Hugo. Vocês são as pessoas que Deus colocou em minha vida para me apoiar e impulsionar a ser a melhor versão de mim mesmo.

Quero também agradecer ao meu tio, Alberto Lachos, e à minha tia, Ana Matute, pelo apoio e pelo exemplo inspirador como pessoas, que me motivam a continuar me aprimorando a cada dia, tanto pessoal quanto profissionalmente.

Ao meu orientador, Jorge Luis Bazán Guzmán, minha extensa gratidão por ser a pessoa essencial e inspiradora ao longo deste processo. Suas orientações e ensinamentos valiosos não apenas me guiaram no desenvolvimento deste projeto, mas também moldaram significativamente minha compreensão e apreciação pela área estatística. Além disso, sua paciência, disponibilidade e compromisso em orientar-me durante os desafios deste caminho acadêmico são verdadeiramente inestimáveis, dado que o conhecimento que adquiri nestes últimos anos foi extraordinário e imensamente essencial.

Finalmente, quero agradecer à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/MEC/Governo do Brasil pelo apoio financeiro concedido durante meu curso de Pós-Graduação em Estatística (PIPGEs).

*“O mais sábio é aquele  
que sabe que não sabe nada.”  
(Sócrates)*

# RESUMO

LACHOS OLIVARES, V. E. **Uma Abordagem Estatística para a Análise dos Resultados das Eleições Presidenciais**. 2024. 87 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Os dados multipartidários possuem características que os tornam dados composicionais tais como uma soma constante de componentes e um espaço limitado conhecido como simplex. Assim, o propósito do trabalho é desenvolver uma metodologia para analisar os dados multipartidários de eleições eleitorais considerando sua natureza restrita. Nesse contexto, a metodologia proposta consiste em 8 etapas: inicialmente, coletamos os dados multipartidários e os transformamos em dados composicionais. Em seguida, aplicamos a transformação de razões logarítmicas, removendo as restrições inerentes de dados composicionais.

Posteriormente, empregamos a análise de componentes principais (ACP) para reduzir a dimensionalidade e identificar os principais componentes que retêm a maior parte da variação dos dados. Essas componentes são analisadas com base em duas métricas importantes: cargas e escores. Dado que os escores possuem diferente variabilidade nas componentes, eles são transformados entre valores zero e um.

Subsequentemente, propomos o modelo de regressão Beta considerando os escores como variável resposta, e os indicadores de desenvolvimento humano como as variáveis explicativas. A metodologia é aplicada nos dados multipartidários das eleições do primeiro turno no Peru em 2021 e no Brasil em 2022, permitindo-nos identificar os principais componentes e que covariáveis (saúde, educação e renda) estão relacionadas diretamente aos votos em diferentes regiões e estados.

Finalmente, considerando que os dados das eleições presidenciais do Peru em 2021 apresentam duas variáveis resposta, propomos um modelo de regressão bivariado via cópulas e analisar a estrutura dependência entre essas variáveis.

**Palavras-chave:** Análise de componentes principais, Transformação de razão logarítmica, Dados composicionais, Escores, Cargas, Modelo de regressão Beta, Cópulas..

# ABSTRACT

LACHOS OLIVARES, V. E. **An statistical approach for analysis of presidential elections results**. 2024. 87 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Multiparty data has characteristics that make it compositional data such as a constant sum of components and a limited space known as simplex. Thus, the purpose of the work is to develop a methodology to analyze multi-party data from electoral elections considering their restricted nature. In this context, the proposed methodology consists of 8 steps: initially, we collect multi-party data and transform it into compositional data. Then, we apply the log-ratio transformation, removing the inherent constraints of compositional data.

Next, we employ principal component analysis (PCA) to reduce dimensionality and identify the principal components that retain most of the variation in the data. These components are analyzed based on two important metrics: loadings and scores. Given that the scores have different variability in the components, they are transformed between values of zero and one.

Subsequently, we propose the Beta regression model considering the scores as the response variable, and the human development indicators as the explanatory variables. The methodology is applied to multiparty data from the first round elections in Peru in 2021 and Brazil in 2022, allowing us to identify the main components and which covariates (health, education and income) are directly related to votes in different regions and states.

Finally, considering that data from presidential elections of Peru 2021 with two response variables, we propose a bivariate regression model via copulas and analyze the dependence structure between these variables.

**Keywords:** Principal Component Analysis, Log-ratio transformation, Compositional Data, Scores, Loadings, Beta Regression Model, Copulas..

# LISTA DE ILUSTRAÇÕES

---

---

|                                                                                                                                 |    |
|---------------------------------------------------------------------------------------------------------------------------------|----|
| Figura 1 – O simplex de duas partidos políticos nas eleições presidenciais no segundo turno . . . . .                           | 18 |
| Figura 2 – diagrama geral da metodologia . . . . .                                                                              | 31 |
| Figura 3 – Diagrama da metodologia aplicada nas eleições peruanas de 2021 . . . . .                                             | 42 |
| Figura 4 – Biplot da ACP dos dados (98,3% da variância obtida pelas duas primeiras componentes) . . . . .                       | 45 |
| Figura 5 – Biplot da ACP das proporções dos dados(77,3% da variância obtida pelas duas primeiras componentes) . . . . .         | 46 |
| Figura 6 – Biplot da ACP da transformação clr dos dados (81.06% da variância obtida pelas duas primeiras componentes) . . . . . | 47 |
| Figura 7 – Densidades dos escores transformados $z_1$ e $z_2$ . . . . .                                                         | 49 |
| Figura 8 – Diagnóstico de resíduos para $z_1$ no Peru . . . . .                                                                 | 54 |
| Figura 9 – Diagnóstico de resíduos para $z_2$ no Peru . . . . .                                                                 | 54 |
| Figura 10 – Diagrama da metodologia nas eleições Brasileiras 2022 . . . . .                                                     | 59 |
| Figura 11 – Biplot da ACP para dados transformados(84% da variância obtida pelas duas primeiras componentes) . . . . .          | 61 |
| Figura 12 – Histograma do escore transformado $z_1$ . . . . .                                                                   | 63 |
| Figura 13 – Diagnóstico de resíduos para $z_1$ no Brasil . . . . .                                                              | 65 |

# LISTA DE CÓDIGOS-FONTE

---

---

|                                                                                     |    |
|-------------------------------------------------------------------------------------|----|
| Código-fonte 1 – Código do modelo em R das eleições peruanas . . . . .              | 78 |
| Código-fonte 2 – Código do modelo em R das eleições brasileiras . . . . .           | 82 |
| Código-fonte 3 – Código do modelo em R das eleições Peruanas utilizando cópulas . . | 86 |

# LISTA DE TABELAS

---

---

|                                                                                                                                                                                |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Tabela 1 – Cópulas bivariadas Elipticas . . . . .                                                                                                                              | 36 |
| Tabela 2 – Principais famílias de cópulas arquimidianas . . . . .                                                                                                              | 37 |
| Tabela 3 – Dados das eleições peruanas de 2021 . . . . .                                                                                                                       | 40 |
| Tabela 4 – Matriz de correlação considerando o número de votos no Peru . . . . .                                                                                               | 43 |
| Tabela 5 – Matriz de correlação considerando as proporções de votos no Peru . . . . .                                                                                          | 44 |
| Tabela 6 – Matriz de correlação considerando as transformações <i>clr</i> dos votos no Peru . . . . .                                                                          | 44 |
| Tabela 7 – Cargas da análise de componentes principais dos dados . . . . .                                                                                                     | 44 |
| Tabela 8 – Cargas da análise de componentes principais das proporções dos dados . . . . .                                                                                      | 45 |
| Tabela 9 – Cargas da análise de componentes principais da transformação <i>clr</i> dos dados . . . . .                                                                         | 47 |
| Tabela 10 – Escores transformados das componentes $x_1$ e $x_2$ . . . . .                                                                                                      | 49 |
| Tabela 11 – Indicadores de desenvolvimento humano do Peru . . . . .                                                                                                            | 50 |
| Tabela 12 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente $x_1$ . . . . .                                                    | 51 |
| Tabela 13 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente $x_1$ sem a variável Educação 2 . . . . .                          | 52 |
| Tabela 14 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente $x_1$ sem as variáveis Educação 2 e Saúde . . . . .                | 52 |
| Tabela 15 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente $x_2$ . . . . .                                                    | 53 |
| Tabela 16 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente $x_2$ sem a variável Saúde . . . . .                               | 53 |
| Tabela 17 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente $x_2$ considerando unicamente covariáveis significativas . . . . . | 53 |
| Tabela 18 – Dados das eleições de Brasil 2022 . . . . .                                                                                                                        | 57 |
| Tabela 19 – Matriz de correlação considerando as transformações <i>clr</i> dos votos nas eleições no Brasil em 2022 . . . . .                                                  | 60 |
| Tabela 20 – Cargas da análise de componentes principais da transformação <i>clr</i> dos dados de Brasil . . . . .                                                              | 60 |
| Tabela 21 – Escores transformados da componente $x_1$ . . . . .                                                                                                                | 62 |
| Tabela 22 – Indicadores de desenvolvimento humano do Brasil . . . . .                                                                                                          | 64 |
| Tabela 23 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente $x_1$ no Brasil . . . . .                                          | 64 |

|                                                                                                         |    |
|---------------------------------------------------------------------------------------------------------|----|
| Tabela 24 – Estimativas dos parâmetros do modelo de regressão Beta sem a variável Longevidade . . . . . | 65 |
| Tabela 25 – Critério de seleção de modelos . . . . .                                                    | 70 |
| Tabela 26 – Estimativas dos parâmetros para preferência pelos partidos progressistas . . .              | 70 |
| Tabela 27 – Estimativas dos parâmetros para preferência pelos partidos tradicionais . . .               | 70 |
| Tabela 28 – Estimativa de $\theta$ . . . . .                                                            | 71 |

# SUMÁRIO

---

---

|       |                                                                                                          |    |
|-------|----------------------------------------------------------------------------------------------------------|----|
| 1     | INTRODUÇÃO . . . . .                                                                                     | 15 |
| 2     | PRELIMINARES . . . . .                                                                                   | 17 |
| 2.1   | Dados composicionais em dados com múltiplos partidos . . . . .                                           | 17 |
| 2.2   | Transformação de razão logarítmica centrada dos dados composicionais . . . . .                           | 19 |
| 2.3   | Análise de componentes principais dos dados . . . . .                                                    | 20 |
| 2.4   | Transformação dos escores . . . . .                                                                      | 22 |
| 2.5   | Modelo de Regressão Beta . . . . .                                                                       | 23 |
| 2.5.1 | <i>Estimação dos parâmetros</i> . . . . .                                                                | 24 |
| 2.5.2 | <i>Teste de hipótese</i> . . . . .                                                                       | 26 |
| 2.5.3 | <i>Medidas de diagnóstico</i> . . . . .                                                                  | 27 |
| 2.6   | Proposta de metodologia . . . . .                                                                        | 29 |
| 2.7   | Revisão de cópulas . . . . .                                                                             | 31 |
| 3     | UMA ABORDAGEM ESTATÍSTICA PARA A ANÁLISE DAS ELEIÇÕES PRESIDENCIAIS DO PERU 2021 . . . . .               | 39 |
| 3.1   | Introdução . . . . .                                                                                     | 39 |
| 3.2   | Metodologia . . . . .                                                                                    | 41 |
| 3.3   | Resultados . . . . .                                                                                     | 43 |
| 3.3.1 | <i>Matrizes de correlação das eleições Peruanas de 2021</i> . . . . .                                    | 43 |
| 3.3.2 | <i>Identificando as principais componentes das eleições Peruanas</i> . . . . .                           | 43 |
| 3.3.3 | <i>Explicando as componentes principais <math>x_1</math> e <math>x_2</math></i> . . . . .                | 48 |
| 3.3.4 | <i>Transformando os escores das componentes principais <math>x_1</math> e <math>x_2</math></i> . . . . . | 48 |
| 3.3.5 | <i>Proposta do modelo de regressão Beta</i> . . . . .                                                    | 48 |
| 3.3.6 | <i>Diagnóstico dos resíduos</i> . . . . .                                                                | 53 |
| 3.4   | Discussão final . . . . .                                                                                | 54 |
| 4     | UMA ABORDAGEM ESTATÍSTICA PARA A ANÁLISE DAS ELEIÇÕES PRESIDENCIAIS DO BRASIL 2022 . . . . .             | 56 |
| 4.1   | Introdução . . . . .                                                                                     | 56 |
| 4.2   | Metodologia . . . . .                                                                                    | 57 |
| 4.3   | Resultados . . . . .                                                                                     | 59 |
| 4.3.1 | <i>Matriz de correlação das eleições Brasileiras de 2022</i> . . . . .                                   | 59 |

|       |                                                                                    |    |
|-------|------------------------------------------------------------------------------------|----|
| 4.3.2 | <i>Identificando as principais componentes das eleições Brasileiras</i> . . .      | 60 |
| 4.3.3 | <i>Explicando a componente principal <math>x_1</math></i> . . . . .                | 61 |
| 4.3.4 | <i>Transformando os escores da componente principal <math>x_1</math></i> . . . . . | 62 |
| 4.3.5 | <i>Proposta do modelo de regressão Beta</i> . . . . .                              | 63 |
| 4.3.6 | <i>Diagnóstico dos resíduos</i> . . . . .                                          | 65 |
| 4.4   | Discussão final . . . . .                                                          | 65 |
| 5     | <b>MODELO DE REGRESSÃO BIVARIADO VIA CÓPULAS: ELEIÇÕES DE PERU</b> . . . . .       | 67 |
| 5.1   | Introdução . . . . .                                                               | 67 |
| 5.2   | Metodologia . . . . .                                                              | 68 |
| 5.3   | Resultados . . . . .                                                               | 69 |
| 5.4   | Discussão final . . . . .                                                          | 71 |
| 6     | <b>CONSIDERAÇÕES FINAIS</b> . . . . .                                              | 72 |
|       | <b>REFERÊNCIAS</b> . . . . .                                                       | 74 |
|       | <b>APÊNDICE A</b> <i>CÓDIGOS R</i> . . . . .                                       | 78 |

---

## INTRODUÇÃO

---

Os dados eleitorais referem-se às informações associadas a um processo eleitoral, no qual os eleitores escolhem representantes para uma circunscrição específica, como estados, regiões ou distritos. Esse processo eleitoral é fundamental para os princípios da democracia, revelando as preferências dos eleitores em relação a partidos políticos e a confiabilidade desses dados é crucial para garantir a legitimidade do governo, assegurar a integridade do processo eletivo, e promover o desenvolvimento e a estabilidade socioeconômica.

Quando analisamos os dados eleitorais, no entanto, é possível identificar um problema, uma vez que esses dados são composicionais. Dados composicionais significam que podem ser representados por proporções, frações ou porcentagens, fornecendo informações relativas em vez de absolutas. Alguns exemplos adicionais de dados composicionais ([BAZAN; SULMONT; CALDERÓN, 2012](#)) incluem a distribuição de gastos familiares, a composição do portfólio de investimentos, o uso diário do tempo em diversas atividades e a distribuição de vendas em diferentes regiões. Esses dados não são exclusivos de uma única área e surgem em ciências políticas, geoquímica, genética, ecologia, entre outras disciplinas. Além disso, os dados composicionais possuem características de não-negatividade e uma soma constante, tornando as variáveis envolvidas dependentes entre si, e conforme citado por [Mosimann \(1962\)](#), há uma correlação espúria gerada por variáveis que representam partes de um todo, assim também esses dados apresentam restrições no espaço dimensional.

Por outro lado, os dados eleitorais são considerados dados multivariados, e para melhorar a interpretação da distribuição dos votos, devemos aplicar a ferramenta estatística de análise de componentes principais, visando reduzir a dimensionalidade. No entanto, as características inerentes e as restrições associadas aos dados composicionais dificultam a aplicação efetiva desse método para redução de dimensionalidade nos dados eleitorais.

Consequentemente, [Aitchison \(1982\)](#) estabeleceu os fundamentos teóricos e metodológicos para a análise estatística de dados composicionais, propondo uma transformação de razão

logarítmica centrada nos dados. Posteriormente, aplicou-se a análise de componentes principais para identificar os principais componentes que explicam a maior parte da variação nos dados eleitorais.

Com as circunscrições agora representadas em um novo espaço dimensional de componentes, observamos uma nova disposição espacial. O indicador, denominado *escore*, obtido através da análise de componentes principais, oferece uma interpretação dessa nova disposição em que as circunscrições são então organizadas de maneira que possibilita a identificação de variáveis latentes por meio dos *escores*. Por tanto, é necessário identificar as variáveis explicativas associadas às circunscrições que explicam por que elas ocupam determinadas posições, assim como essas variáveis explicativas podem explicar as variáveis latentes associadas aos componentes principais.

Dentre as pesquisas que exploraram os dados eleitorais usando análise de dados composicionais, destacam-se os trabalhos [Rodrigues e Lima \(2009\)](#) e [Bazan, Sulmont e Calderón \(2012\)](#). No primeiro caso, foi apresentada uma análise das eleições da União Europeia usando análise de componentes principais, enquanto o segundo considerou dados das eleições peruanas de 2011.

Dessa forma, nesta dissertação propomos uma metodologia para analisar os resultados das eleições presidenciais no Peru em 2021 e nas eleições brasileiras em 2022. A proposta é modelar as principais componentes, considerando variáveis ligadas às respectivas circunscrições. Isso será feito aplicando modelos de regressão e agregando variáveis explicativas relacionadas ao comportamento político da sociedade em questão. Além disso, é importante notar que pode haver mais de uma componente relacionada às variáveis, razão pela qual propomos o uso de modelos de regressão bivariada via cópulas para identificar a provável dependência entre as componentes na presença de variáveis explicativas.

A dissertação está organizado da seguinte forma: no [Capítulo 2](#), apresentamos os conceitos preliminares necessários para compreender a metodologia, incluindo a definição dos dados composicionais, a transformação de razões logarítmicas em dados composicionais e a análise de componentes principais, onde obtemos duas métricas importantes: as cargas e os *escores*. Devido à variabilidade dos *escores*, exploramos a necessidade de uma transformação para que eles estejam dentro do intervalo entre zero e um. Além disso, propomos o modelo de Regressão Beta, considerando os *escores* como variável resposta. No [Capítulo 3](#), aplicamos a metodologia para analisar os votos das eleições peruanas de 2021. No [Capítulo 4](#), aplicamos a metodologia para analisar os votos das eleições Brasileiras de 2022. No [Capítulo 5](#) desenvolvemos um modelo bivariado aplicando cópulas nas eleições peruanas de 2021. Por último, no [Capítulo 6](#), resumimos os resultados mais importantes obtidos na análise das eleições peruanas e brasileiras usando o pacote GAMLSS e SemiParBIVProbit para o modelo univariado e bivariado respectivamente.

---

## PRELIMINARES

---

### 2.1 Dados composicionais em dados com múltiplos partidos

Identificamos as características de dados eleitorais multipartidários considerando  $V_{ij}$ , que denota a proporção de votos na circunscrição eleitoral onde a denotação para região, província, estado ou distrito é  $i$  e para os partidos é  $j$ .

Dados com votos multipartidários apresentam duas características principais, que são:

O intervalo de valores para  $V_{ij}$ , onde:

$$V_{ij} \in [0, 1], i = 1, \dots, n \text{ e } j = 1, \dots, p$$

E a soma das proporções de votos para todos os partidos em cada circunscrição é igual a 1, para  $i$  variando de 1 a  $n$ :

$$\sum_{j=1}^p V_{ij} = 1, i = 1, \dots, n.$$

As variáveis que satisfazem as duas características mencionadas geralmente estão em uma região denominada simplex. O simplex é definido da seguinte maneira:

$$\mathbf{V}_i = (V_{i1}, \dots, V_{ip}) : \sum_{j=1}^p V_{ij} \leq 1, i = 1, \dots, n$$

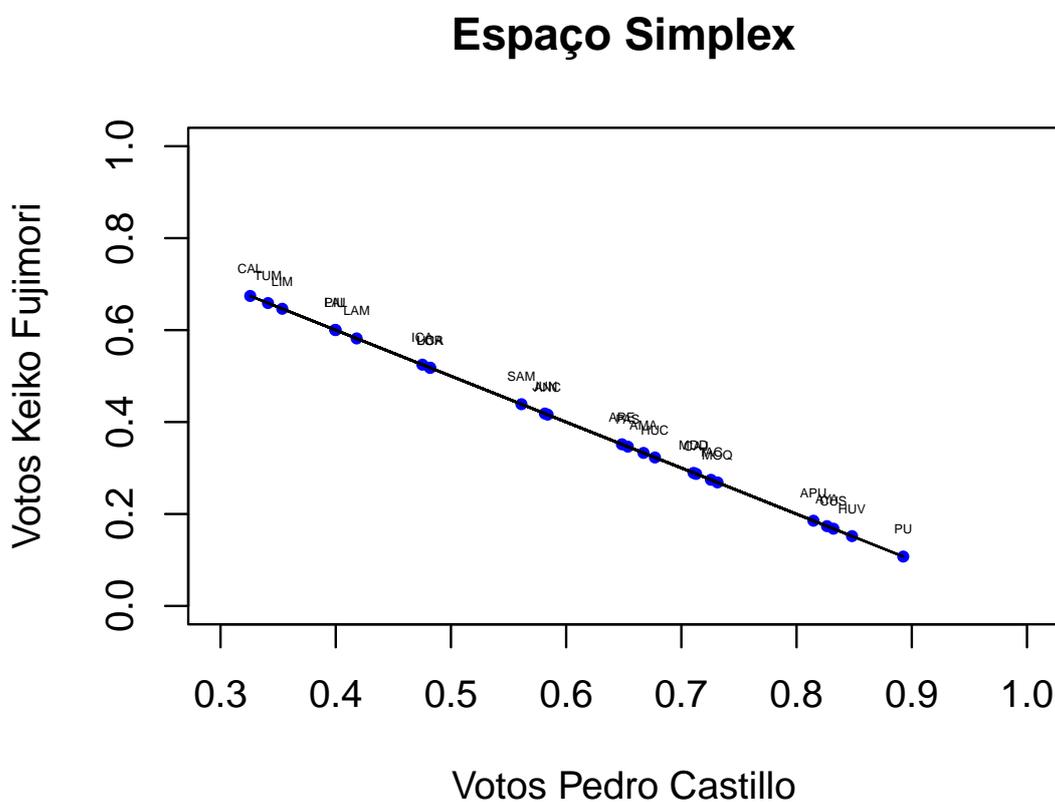
Aqui,  $\mathbf{V}_i$  representa um vector para uma região específica  $i$ , e  $j$  denota os partidos políticos, onde  $p$  indica a dimensão do vector.

Para uma melhor visualização do espaço simplex, a continuação apresentamos os votos eleitorais

do segundo turno das eleições presidenciais de 2021 no Peru, que envolveram dois candidatos: Keiko Fujimori  $V_{iF}$  e Pedro Castillo  $V_{iC}$ . Aqui,  $i$  denota as diferentes 24 regiões e a única província constitucional do Peru, e conforme representado na [Figura 1](#) utilizamos a padronização proposta por [ISO3166 \(1998\)](#) para cada região. Podemos representar facilmente ambas as variáveis como  $V_{iC} + V_{iF} = 1$ , o que nos permite entender as proporções de votos dos candidatos em relação ao espaço simplex.

De acordo com a segunda restrição, todas as proporções de votos das circunscrições estão localizadas em uma linha com limites nos eixo  $X$  e  $Y$ , abaixo ou igual a 1, devido à primeira restrição.

Figura 1 – O simplex de duas partidos políticos nas eleições presidenciais no segundo turno



*Amazonas(AMA), Ancash(ANC), Apurímac(APU), Arequipa(ARE), Ayacucho(AYA), Cajamarca(CAJ), Cusco(CUS), Callao(CAL), Huancavelica(HUV), Huánuco(HUC), Ica(ICA), Junín(JUN), La Libertad(LAL), Lambayeque(LAM), Lima(LIM), Loreto(LOR), Madre de Dios(MDD), Moquegua(MOQ), Pasco(PAS), Piura(PIU), Puno(PUN), San Martín(SAM), Tacna(TAC), Tumbes(TUM), Ucayali(UCA).*

Fonte: Elaborada pelo autor.

A mesma lógica é aplicada a eleições com mais de dois partidos ( $p > 2$ ). No entanto, o gráfico torna-se mais complexo. Para ilustrar, consideramos uma eleição com três partidos políticos, onde  $V_{iF}$  representa os votos para a candidata Keiko Fujimori,  $V_{iC}$  representa os votos para o candidato Pedro Castillo e  $V_{iO}$  representa os votos acumulados para outros candidatos. Nesse caso, obtemos um tetraedro no espaço tridimensional, representando as proporções de

votos para os diferentes candidatos nas circunscrições eleitorais.

## 2.2 Transformação de razão logarítmica centrada dos dados composicionais

Uma metodologia adequada para a análise de dados composicionais em eleições multipartidárias deve levar em consideração alguns princípios e características do espaço simplex. A ideia central é que dados composicionais e votos dos partidos políticos fornecem apenas informações sobre a magnitude relativa das partes, não justificando interpretações que envolvam magnitudes absolutas. Assim, assumimos que o valor constante da soma das partes é irrelevante.

Ao reconhecer a importância das proporções em dados composicionais, surge a pergunta sobre que tipo de proporções estamos considerando. Nesse contexto, o fundamento dessa metodologia foi proposto por [Aitchison \(1982\)](#), consistindo em uma transformação dos dados composicionais, originalmente no simplex  $\mathbf{V}_i$ , em um vetor que envolve proporções entre as partes e é definido em um espaço real. Quando essa transformação é bijetiva, estabelece uma correspondência direta entre a composição no simplex e uma representação real em forma de vetor. Isso significa que qualquer problema relacionado a dados composicionais pode ser expresso em vetores transformados, permitindo-nos resolver essas questões utilizando técnicas multivariadas em espaços reais.

Consideremos, por exemplo,  $n$  circunscrições eleitorais com proporções de votos  $V_{ij}$ , onde  $i = 1, \dots, n$  e  $j = 1, \dots, p$ . Inicialmente, assumimos que essas proporções são composicionais. No caso em que não são composicionais, podemos aplicar a transformação sugerida por [Egozcue et al. \(2003\)](#). No entanto, neste trabalho, adotamos a proposta de [Aitchison \(1982\)](#), conhecida como transformação de razão logarítmica centrada (clr) em  $R^p$ , e definida como:

$$clr(\mathbf{V}_i^*) = \left[ \log \left( \frac{V_{i1}}{g_i} \right), \dots, \log \left( \frac{V_{ip}}{g_i} \right) \right]^\top,$$

onde  $g_i$  é a média geométrica das proporções na  $i$ -ésima circunscrição eleitoral, ou seja,

$$g_i = \left[ \prod_{j=1}^p V_{ij} \right]^{\frac{1}{p}} \Rightarrow \log(g_i) = \frac{\sum_{j=1}^p \log V_{ij}}{p},$$

outra forma equivalente de representar a equação é:

$$clr(\mathbf{V}_i^*) = [\log(V_{i1}) - \log(g_i), \dots, \log(V_{ip}) - \log(g_i)]^\top$$

Além disso, observamos que os valores transformados das proporções de votos são denotados como  $V_{ij}^* = \log \left( \frac{V_{ij}}{g_i} \right)$  e temos uma propriedade fundamental  $\sum_{j=1}^p V_{ij}^* = 0$  para  $i = 1, \dots, n$ . O fato de aplicarmos logaritmos nas proporções se trata de uma conveniência matemática. O logaritmo

de quocientes é mais manejável e permite algumas propriedades simples. Para ilustrar, não há um relacionamento direto entre as variâncias  $Var(\frac{V_{im}}{V_{in}})$  e  $Var(\log(\frac{V_{im}}{V_{in}})) = Var(\log(\frac{V_{im}}{V_{in}}))$  porque o logaritmo transforma as razões em diferenças que são linearmente relacionadas, simplificando os cálculos e não afetando a variância, pois:

$$Var(\log(\frac{V_{im}}{V_{in}})) = Var(-\log(\frac{V_{in}}{V_{im}})) = Var(\log(\frac{V_{in}}{V_{im}})).$$

Existem muitas outras propostas de transformações, como a razão logarítmica aditiva (*alr*) e a razão logarítmica isométrica (*ilr*), que apresentam características desejáveis. No entanto, a transformação *alr* está vinculada à assimetria e, em sua definição, é uma transformação não isométrica para o simplex. Por outro lado, a transformação *ilr* aumenta em sua complexidade, o que dificulta a interpretação das coordenadas (EGOZCUE *et al.*, 2003). Em contraste, a transformação *clr* está vinculada ao contexto não paramétrico e é possível especificar a distância de Aitchison em termos de distância euclidiana com os vetores transformados *clr*,

Além disso, a transformação *clr* é simétrica e isométrica, mas a imagem  $\mathbf{V}_i$  permanece restrita ao subespaço  $R^p$  e a matriz de covariância dos dados transformados *clr* é singular. Após a aplicação de *clr* em  $p$  variáveis, a soma das  $p$  variáveis é zero para cada circunscrição  $i$ . Se alguns dos dados originais forem zero, a transformação *clr* não pode ser realizada da mesma maneira. No entanto, podemos aplicar um método de substituição onde todos os elementos zero são substituídos por  $0 + \varepsilon$ , onde  $\varepsilon$  é um pequeno valor, a fim de não afetar os resultados na análise.

## 2.3 Análise de componentes principais dos dados

A análise de componentes principais, proposta por Pearson (1901), consiste na redução da dimensionalidade de um conjunto de dados para obter maior interpretabilidade e uma melhor visualização geométrica dos dados no espaço. Consequentemente, as principais componentes são obtidas por meio da matriz de variância-covariância, e explicam a maioria da variância dos dados. Como exemplo, consideramos o vetor aleatório denotado como  $\mathbf{V}^* = (V_1^*, V_2^*, \dots, V_p^*)^\top$ , onde  $p$  representa o número de variáveis, com matriz de variância-covariância  $\Sigma$  (JOHNSON; WICHERN *et al.*, 2002) denotado como:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1p}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{p1}^2 & \cdots & \sigma_{pp}^2 \end{pmatrix}$$

em que seus autovalores e autovetores são  $\boldsymbol{\lambda}_j = (\lambda_1, \dots, \lambda_p)$  e  $\mathbf{e}_j = (\mathbf{e}_1, \dots, \mathbf{e}_p)$  respectivamente. Então, consideramos as seguintes combinações lineares:

$$\begin{aligned}
X_1 &= \mathbf{a}_1 \mathbf{V}^* = a_{11}V_1^* + a_{12}V_2^* + \dots + a_{1p}V_p^* \\
X_2 &= \mathbf{a}_2 \mathbf{V}^* = a_{21}V_1^* + a_{22}V_2^* + \dots + a_{2p}V_p^* \\
&\vdots \\
X_p &= \mathbf{a}_p \mathbf{V}^* = a_{p1}V_1^* + a_{p2}V_2^* + \dots + a_{pp}V_p^*
\end{aligned}$$

em que,

$$\begin{aligned}
\text{var}(X_j) &= \mathbf{a}_j^\top \mathbf{\Sigma} \mathbf{a}_j = \lambda_j, & j = 1, \dots, p \\
\text{cov}(X_j, X_k) &= \mathbf{a}_j^\top \mathbf{\Sigma} \mathbf{a}_k = 0, & j \neq k
\end{aligned}$$

As principais componentes são as combinações lineares não correlacionadas  $Y_1, \dots, Y_p$  que maximizam a variância. Seguidamente, as primeiras componentes explicam a maior quantidade de variância nos dados, enquanto as outras componentes subsequentes explicam uma quantidade menor dessa variância dos dados. Isso ocorre devido à falta de correlação entre as componentes e à sua ortogonalidade. Posteriormente, observamos que a primeira componente apresenta a combinação linear com a variância máxima, ou seja,  $\text{var}(Y_1) = \mathbf{a}_1^\top \mathbf{\Sigma} \mathbf{a}_1$ , onde  $\mathbf{a}_1$  é o vetor de pesos associado à primeira componente e  $\mathbf{\Sigma}$  é a matriz de covariância dos dados. Da mesma forma, a segunda componente apresenta a combinação linear com a máxima variância residual, isto é,  $\text{var}(Y_2) = \mathbf{a}_2^\top \mathbf{\Sigma} \mathbf{a}_2$ , e assim sucessivamente para as demais componentes.

Assim, seja uma matriz de variância-covariância  $\mathbf{\Sigma}$  em que  $\mathbf{a}_j = \mathbf{e}_j$  de um vetor aleatório  $\mathbf{V}^* = (V_1^*, V_2^*, \dots, V_p^*)^\top$ . Assim, a  $i$ -ésima componente principal é dada por,

$$X_j = \mathbf{e}_j^\top \mathbf{V}_j^* = e_{j1}V_1^* + e_{j2}V_2^* + \dots + e_{jp}V_p^*, \quad j = 1, \dots, p$$

onde,  $X_j$  caracteriza a  $j$ -ésima componente principal, representando um novo espaço dimensional de componentes. Portanto, temos que,

$$\begin{aligned}
\text{var}(X_j) &= \mathbf{e}_j^\top \mathbf{\Sigma} \mathbf{e}_j = \lambda_j, & j = 1, \dots, p \\
\text{cov}(X_j, X_k) &= \mathbf{e}_j^\top \mathbf{\Sigma} \mathbf{e}_k = 0, & j \neq k
\end{aligned}$$

em que a primeira igualdade indica que a variância de cada componente  $X_j$  é o autovalor  $\lambda_j$ , respectivamente. Na segunda igualdade, a covariância entre duas componentes diferentes distintas é zero, indicando que as componentes não estão correlacionadas.

Como [Rodrigues e Lima \(2009\)](#) observou, entre os abordagens possíveis para a Análise de Componentes Principais (ACP) em relação aos dados eleitorais consideramos: a frequência

dos dados eleitorais (número de votos), dados eleitorais transformados em proporções, e dados proporcionais transformados em razões logarítmicas.

O primeiro é conhecido como análise de componentes principais direta e podemos perceber que a soma dos votos é diferente para cada província ou região. Além disso, se colocarmos os votos no espaço multidimensional e aplicarmos PCA, a análise mostra certa curvatura a qual é incompatível com a hipótese de linearidade de análise de principais componentes e leva a resultados insatisfatórios (RODRIGUES; LIMA, 2009). No segundo caso, a Análise de Componentes Principais nas proporções é aplicada em dados onde os votos eleitorais são substituídos por uma distribuição relativa (proporções) do total por cada região, onde os diagramas dispersos correspondentes aos pares devem ser  $(\mathbf{e}'_i \mathbf{x}_r, \mathbf{e}'_j \mathbf{x}_r) : r = 1, \dots, n$ , e  $\mathbf{x}_r$  é o vetor com os votos eleitorais na região  $r$  (RODRIGUES; LIMA, 2009). Além disso, a matriz de variância-covariância associada aos votos proporcionais é  $\Gamma = [\text{cov}(\mathbf{x}_i, \mathbf{x}_j)]$  com autovalores positivos  $\lambda_1 > \lambda_2 > \dots > \lambda_p > 0$  e autovetores correspondentes  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ . Se os pontos no gráfico disperso apresentam uma estrutura elíptica homogênea, a Análise de Componentes Principais para proporções é uma solução adequada. No entanto, se observamos padrões curvos, o método produz resultados inadequados. Por último, a Análise de Componentes Principais das transformações de razões logarítmicas centradas (clr) dos dados indica uma alternativa melhor na análise de dados, pois corrige a restrição de soma constante e reduz a dependência entre as variáveis.

## 2.4 Transformação dos escores

Ao aplicar a Análise de Componentes Principais (ACP), obtemos duas medidas essenciais: cargas e escores. As cargas indicam as relações das variáveis com as componentes, enquanto os escores representam as posições das observações no novo espaço dimensional das componentes.

No entanto, os escores  $x_{ij}$  em que  $i = 1, \dots, n$  e  $j = 1, \dots, p$ , têm variabilidades diferentes em cada componente, o que dificulta a interpretação. Para abordar essa questão, sugere-se realizar uma transformação monotônica nos escores, a fim de padronizá-los em uma escala comum de  $(0, 1)$  para cada componente. Essa transformação é efetuada utilizando a seguinte fórmula:

$$z_{ij} = \frac{x_{ij} - \min(\mathbf{X}_j)}{\max(\mathbf{X}_j) - \min(\mathbf{X}_j)}$$

Nesta equação,  $z_{ij}$  representa os escores transformados para cada componente  $j$ , e  $\mathbf{X}_j = (x_{1j}, \dots, x_{nj})^T$  é o vetor  $n$ -dimensional de escores não transformados da componente  $j$ . A fórmula leva em consideração o máximo e mínimo dos escores em cada  $\mathbf{X}_j$ , e assegura que os escores transformados sejam comparáveis de forma consistente em todas as componentes, tornando a análise mais acessível e fácil de compreender.

## 2.5 Modelo de Regressão Beta

A distribuição Beta é conhecida por sua flexibilidade e utilidade ao modelar frações, porcentagens, proporções ou variáveis com valores entre 0 e 1. Uma variável aleatória  $Z$  tem distribuição Beta se sua função de densidade é dada por:

$$f(z|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1} (1-z)^{(\beta-1)}$$

em que,  $\alpha > 0$  e  $\beta > 0$  são os parâmetros de forma, e  $\Gamma$  denota a função Gamma, isto é,  $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ . A média e a variância de  $Z$  são, respectivamente,

$$E(z) = \frac{\alpha}{\alpha + \beta}$$

$$var(z) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Além disso, distribuição uniforme é um caso particular quando  $\alpha = \beta = 1$ . Em modelos de regressão, é conveniente utilizar uma parametrização diferente, onde é comum definir a média que compreenda o parâmetro de precisão ou dispersão. Assim, em modelos de regressão, a distribuição beta é reparametrizada, ou seja,  $\mu = \frac{\alpha}{\alpha + \beta}$  e  $\phi = \alpha + \beta$ , em que  $\mu_i$  e  $\phi$  são a média e o parâmetro de precisão respectivamente. Nesse contexto, a função de densidade pode ser escrita como:

$$f(z|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} z^{\mu\phi-1} (1-z)^{(1-\mu)\phi-1} \quad (2.1)$$

em que,  $\mu \in (0, 1)$  e  $\phi > 0$ . Portanto,

$$E(z) = \mu$$

$$var(z) = \frac{\mu(1-\mu)}{(1+\phi)}.$$

Sejam  $Z_1, \dots, Z_n$  variáveis aleatórias independentes, cada uma seguindo uma distribuição Beta mostrado na [Equação 2.1](#) com média  $\mu_i$  e parâmetro de precisão  $\phi$ . O modelo de regressão (FERRARI; CRIBARI-NETO, 2004) é construído considerando o componente aleatório:

$$Z_i \sim Beta(\mu_i, \phi)$$

e com a média  $\mu_i$  descrita pelo componente sistemático:

$$g(\mu_i) = \mathbf{U}_i^\top \boldsymbol{\beta}$$

Aqui,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^\top$  representa o vetor de parâmetros desconhecidos ( $\boldsymbol{\beta} \in \mathbb{R}^k$ ), e  $\mathbf{U}_i = (u_{i1}, \dots, u_{ik})$  são observações em  $k$  covariáveis ( $k < n$ ) que assumem valores fixos e conhecidos. Finalmente,  $g(\cdot)$  é uma função de ligação estritamente monótona e duas vezes diferenciável, mapeando valores de  $(0, 1)$ . Algumas funções de link úteis são: logito  $\log(\frac{\mu}{1-\mu})$ , probito  $\phi^{-1}(\mu)$ , complemento log-log  $\log\{-\log(1-\mu)\}$ , log-log  $-\log\{-\log(\mu)\}$  e cauchy  $\tan\{\pi(\mu - 0.5)\}$ .

### 2.5.1 Estimação dos parâmetros

Identificados que parâmetros estão presentes na distribuição, o processo de estimação dos parâmetros é mediante o método de máxima verossimilhança, proposto por Fisher (1922) que consiste em estimar os parâmetros que assumem a distribuição beta, considerando os  $n$  dados observados. A função de verossimilhança é definido como:

$$L(\boldsymbol{\beta}, \phi; z_i) = \prod_{i=1}^n f(z_i; \boldsymbol{\beta}, \phi)$$

Para facilitar a estimação, aplicamos o logaritmo, também conhecido como a função de log-verossimilhança, baseado em  $n$  observações independentes.

$$l(\boldsymbol{\beta}, \phi; z_i) = \log \left( \prod_{i=1}^n f(z_i; \boldsymbol{\beta}, \phi) \right) = \sum_{i=1}^n l_i(\mu_i, \phi; z_i)$$

em que,

$$l_i(\mu_i, \phi; z_i) = \log \Gamma(\phi) - \log \Gamma(\mu_i \phi) - \log \Gamma((1 - \mu_i) \phi) + (\mu_i \phi - 1) \log z_i + ((1 - \mu_i) \phi - 1) \log(1 - z_i).$$

A função escore do modelo é obtida derivando a log-verossimilhança em relação aos parâmetros desconhecidos, dada como  $(\mathcal{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \phi), \mathcal{U}_{\phi}(\boldsymbol{\beta}, \phi))^\top$  em que  $\mathcal{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \phi)$  é a função escore para o parâmetro  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)$  relacionado com a média  $\mu$  e  $\mathcal{U}_{\phi}(\boldsymbol{\beta}, \phi)$  é a função escore para o parâmetro de precisão  $\phi$ .

As componentes do vetor escore de  $\boldsymbol{\beta}$  são dadas por,

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_k} = \sum_{i=1}^n \frac{\partial l_i(\mu_i, \phi)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \quad k = 1, \dots, q \quad (2.2)$$

com,

$$\frac{\partial l_i(\mu_i, \phi)}{\partial \mu_i} = \phi_i \left[ \log \left( \frac{z_i}{1 - z_i} \right) - \psi(\mu_i \phi) - \psi((1 - \mu_i) \phi) \right] \quad (2.3)$$

em que,  $\psi(\cdot)$  é a função digamma, isto é  $\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ , e  $g'(\mu_i) = \frac{\eta_i}{\mu_i}$ . Portanto, a Equação 2.2 é reduzida a:

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_k} = \sum_{i=1}^n \phi(z_i^* - \mu_i^*) \frac{1}{g'(\mu_i)} \frac{\partial \eta_i}{\partial \beta_k}$$

em que,  $z_i^* = \log\left(\frac{z_i}{1-z_i}\right)$  e  $\mu_i^* = \psi(\mu_i\phi) - \psi((1-\mu_i)\phi)$ . Portanto, a função escore para  $\beta$  em forma matricial é dada por:

$$\mathcal{U}_\beta(\beta, \phi) = \phi \mathbf{U}^\top \mathbf{T}(z^* - \mu^*) \quad (2.4)$$

em que  $\mathbf{U}$  é matriz de covariáveis de ordem  $n \times q$  cuja  $i$ -ésima linha é  $\mathbf{u}_i^\top$ ,  $\mathbf{T} = \text{diag}\left(\frac{1}{g'(\mu_1)}, \dots, \frac{1}{g'(\mu_n)}\right)$ ,  $z^* = (z_1^*, \dots, z_n^*)^\top$  e  $\mu^* = (\mu_1^*, \dots, \mu_n^*)^\top$ . Além disso, a função escore do parâmetro de precisão  $\phi$  é dada por:

$$\mathcal{U}_\phi(\beta, \phi) = \sum_{i=1}^n \{\mu_i(z_i^* - \mu_i^*) + \log(1 - z_i) - \psi((1 - z_i)\phi) + \psi(\phi)\}. \quad (2.5)$$

A próxima etapa é obter a matriz de informação de Fisher. A notação é obtida da seguinte forma. Seja,  $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$ , em que,

$$w_i = \phi \left\{ \psi'(\mu_i\phi) + \psi'((1-\mu_i)\phi) \right\} \left\{ \frac{1}{g'(\mu_n)} \right\}^2,$$

$\mathbf{c} = (c_1, \dots, c_n)^\top$ , em que  $c_i = \phi \left\{ \psi'(\mu_i\phi)\mu_i - \psi'((1-\mu_i)\phi)(1-\mu_i) \right\}$ , onde  $\psi'(\cdot)$  é a função trigamma. Além disso, seja  $\mathbf{D} = \text{diag}\{d_1, \dots, d_n\}$  em que  $d_i = \psi'(\mu_i\phi)\mu_i^2 + \psi'((1-\mu_i)\phi)(1-\mu_i)^2 - \psi'(\phi)$ . Consequentemente, a matriz de Fisher é dada por (BAYER, 2011),

$$\mathbf{K} = K(\beta, \phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix}$$

em que  $K(\beta, \beta) = \phi \mathbf{U}^\top \mathbf{W} \mathbf{U}$ ,  $K(\beta, \phi) = (K(\phi, \beta))^\top = \mathbf{X}^\top \mathbf{T} \mathbf{c}$ , e  $K(\phi, \phi) = \text{tr}(\mathbf{D})$ .

A estimação dos parâmetros,  $\beta_k$  e  $\phi$ , na log-verossimilhança da distribuição beta, é realizada quando Equação 2.4 e Equação 2.5 :

$$\mathcal{U}_\beta(\beta, \phi) = 0$$

$$\mathcal{U}_\phi(\beta, \phi) = 0$$

Esse processo requer um método iterativo, como os algoritmos de Newton-Raphson ou escore de Fisher, para estimar esses parâmetros. A seguir, descrevemos cada um desses processos:

#### Processo iterativo de Newton Raphson

Seja  $\alpha = (\beta^\top, \phi)$ , o vetor de parâmetros, com função escore  $\mathcal{U}(\alpha) = (\mathcal{U}_\beta(\beta, \phi)^\top, \mathcal{U}_\phi(\beta, \phi))^\top$  e dimensão  $(k+1) \times 1$ . O processo consiste em associar aos parâmetros um valor inicial,

$$\mathcal{U}(\alpha) \cong \mathcal{U}(\alpha^{(0)}) + \mathcal{U}'(\alpha^{(0)})(\alpha - \alpha^{(0)})$$

em que  $\mathcal{U}'(\alpha)$  denota a primeira derivada de  $\mathcal{U}(\alpha)$  em relação a  $\alpha^\top$ . Considerando  $\mathcal{U}(\alpha) = 0$  e o processo iterativo, temos que,

$$\alpha^{(m+1)} = \alpha^{(m)} - \mathcal{U}'(\alpha^{(m)})^{-1} \mathcal{U}(\alpha^{(m)}) \quad m = 1, 2, \dots \quad (2.6)$$

*Processo iterativo de escore de Fisher*

Pela lei dos grandes números,  $\mathcal{U}'(\boldsymbol{\alpha}^{(m)})$  converge na matriz de Fisher quando  $n \rightarrow \infty$ . Assim, substituindo na [Equação 2.6](#), temos que,

$$\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} - K'(\boldsymbol{\alpha}^{(m)})\mathcal{U}(\boldsymbol{\alpha}^{(m)}) \quad m = 1, 2, \dots \quad (2.7)$$

**2.5.2 Teste de hipótese**

Quando realizamos uma análise de regressão, buscamos as variáveis independentes que possuem maior significância e influência no modelo. Isso nos permite comparar diferentes modelos e identificar quais variáveis são relevantes para a variável dependente. Nesse contexto, existem diferentes tipos de testes, tais como o teste de razão de verossimilhança, o teste de escore e o teste de Wald ([OLIVEIRA, 2004](#)), dentro da perspectiva do modelo de regressão Beta.

De acordo com certas condições de regularidade, para tamanhos de amostrais grandes, a distribuição conjunta de  $\hat{\boldsymbol{\beta}}$  e  $\hat{\phi}$  ([SEN; SINGER, 1993](#)) é aproximadamente uma distribuição normal q-multivariada, dada por:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\phi} \end{pmatrix} \sim N_q \left( \begin{pmatrix} \boldsymbol{\beta} \\ \phi \end{pmatrix}, \mathbf{K}^{-1} \right)$$

*Teste de razão de verossimilhança*

O teste pode avaliar se o modelo apresenta uma melhora significativa no ajuste. Consideremos a seguinte hipótese:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$$

$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}^{(0)}$$

em que,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^\top$  e  $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \dots, \beta_q^{(0)})^\top$ . A estatística de razão de verossimilhança é dada por:

$$w = 2\{l(\hat{\boldsymbol{\beta}}, \hat{\phi}) - l(\tilde{\boldsymbol{\beta}}, \tilde{\phi})\}$$

em que,  $l(\boldsymbol{\beta}, \phi)$  é o logaritmo natural da função de verossimilhança e  $(\tilde{\boldsymbol{\beta}}, \tilde{\phi})$  é o estimador de máxima verossimilhança restrito pela imposição da hipótese nula. Sob condições de regularidade e sob  $H_0$ ,  $w \xrightarrow{D} \chi_m^2$ .

*Teste escore*

O teste escore é formulado da seguinte maneira:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$$

$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}^{(0)}$$

Considerando  $U_{1\beta}$  como um vetor  $q$  dimensional compreendendo  $q$  elementos da função escore de  $\boldsymbol{\beta}$  e  $\mathbf{K}_{11}^{\beta\beta}$  como uma matriz  $q \times q$  formada pelas  $q$  primeiras linhas e  $q$  primeiras colunas da matriz  $\mathbf{K}^{-1}$ . A estatística é expressada como:

$$w = \tilde{\mathbf{U}}_{1\beta}^\top \tilde{\mathbf{K}}_{11}^{\beta\beta} \tilde{\mathbf{U}}_{1\beta}$$

em que a tilde indica que as quantidades são avaliadas no estimador de máxima verossimilhança restrito. Sob condições de regularidade e sob  $H_0, w \xrightarrow{D} \chi_m^2$ .

### Teste de Wald

O teste de Wald realiza inferência assintótica para os parâmetros do vetor  $\boldsymbol{\beta}$ . Consideremos:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$$

$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}^{(0)}$$

A estatística de teste é definida como:

$$w = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})^\top (\hat{\mathbf{K}}_{11}^{\beta\beta})^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})$$

em que  $\hat{\mathbf{K}}_{11}^{\beta\beta}$  é uma matriz  $q \times q$  formada pelas  $q$  primeiras linhas e  $q$  primeiras colunas da matriz  $\mathbf{K}^{-1}$ , e  $\hat{\mathbf{K}}_{11}^{\beta\beta} = \mathbf{K}_{11}^{\beta\beta}$  considerado no estimador de máxima verossimilhança irrestrito, e  $\hat{\boldsymbol{\beta}}$  é o estimador de máxima verossimilhança de  $\boldsymbol{\beta}$ . Sob condições gerais de regularidade e sob  $H_0, w \xrightarrow{D} \chi_m^2$ .

### 2.5.3 Medidas de diagnóstico

Após ajustar o modelo, é crucial realizar o diagnóstico do modelo proposto para verificar a qualidade do ajuste, identificando possíveis desvios na parte aleatória ( $z_i$ ) e no componente sistemático  $\eta_i$ . Uma medida global da variação pode ser estimada pelo pseudo  $R^2(R_p^2)$ , definido como o quadrado do coeficiente de correlação amostral entre  $\hat{\eta}$  e  $g(z)$ . Note que  $0 \leq R_p^2 \leq 1$ , e um valor mais próximo de 1 indica um melhor ajuste. Outra medida é a discrepância do ajuste, proposta por [Nelder e Wedderburn \(1972\)](#) que é duas vezes a diferença entre o modelo saturado e o modelo postulado. Assim,

$$D(z; \boldsymbol{\mu}, \boldsymbol{\phi}) = \sum_{i=1}^n 2(l_i(\tilde{\boldsymbol{\mu}}_i, \boldsymbol{\phi}) - l_i(\boldsymbol{\mu}_i, \boldsymbol{\phi})),$$

em que  $\tilde{\mu}_i$  é um valor de  $\mu_i$  que resolve  $\frac{\partial l_i}{\partial \mu_i} = 0$ , isto é,  $\phi(z_i^* - \mu_i^*)$ . Quando  $\phi$  é grande,  $\mu_i^* \approx \log\left\{\frac{\mu_i}{1-\mu_i}\right\}$ , e então segue que  $\tilde{\mu}_i \approx z_i$ . Para  $\phi$  conhecido, podemos definir uma medida de discrepância como  $D(z; \tilde{\mu}, \phi)$ , em que  $\tilde{\mu}$  é o estimador de máxima verossimilhança de  $\mu$  sob o modelo pesquisado. Quando  $\phi$  é desconhecido, uma aproximação para essa quantidade é  $D(z; \hat{\mu}, \hat{\phi})$ , denominada comumente de desvio do modelo sob pesquisa. Note que,  $D(z; \hat{\mu}, \hat{\phi}) = \sum_{i=1}^n (r_i^d)^2$

$$r_i^d = \text{sinal}(z_i - \hat{\mu}_i) \{2(l_i(\tilde{\mu}_i, \hat{\phi}) - l_i(\hat{\mu}_i, \hat{\phi}))\}$$

O leverage generalizado, proposto por [Wei et al. \(1998\)](#) é um componente para determinar as observações influentes. Ele é definido como:

$$GL(\tilde{\theta}) = \frac{\partial \tilde{z}}{\partial \tilde{z}^\top},$$

em que,  $\theta$  é um s-vetor tal que  $E(z) = \mu(\theta)$  e  $\tilde{\theta}$  é um estimador de  $\theta$  com  $\tilde{z} = \mu(\tilde{\theta})$ . O elemento  $(i, u)$  de  $GL(\tilde{\theta})$  tal que o leverage generalizado do estimador  $\tilde{\theta}$  em  $(i, u)$  é a taxa de variação em i-ésimo valor predito em relação ao u-ésimo valor resposta. Além disso, os autores notaram que o leverage generalizado  $GL(\tilde{\theta})$  é invariante sob reparametrizações e observações com grande  $GL_{iu}$  são pontos leverage. Seja  $\hat{\theta}$  o estimador de máxima verossimilhança de  $\theta$ , assumindo que existe e é único, e a função de log-verossimilhança de segunda ordem com respeito a  $\theta$  e  $z$ . O leverage generalizado é dado por ([WEI et al., 1998](#)):

$$GL(\theta) = D_\theta \left( \frac{\partial^2 l}{\partial \theta \partial \theta^\top} \right)^{-1} \frac{\partial^2 l}{\partial \theta \partial z^\top}$$

avaliando em  $\hat{\theta}$ , em que  $D_\theta = \frac{\partial \mu}{\partial \theta^\top}$ .

Além disso, considerando um  $\phi$  conhecido, temos o leverage generalizado de  $\beta$  ([FERRARI; CRIBARI-NETO, 2004](#)) dado por,

$$GL(\beta) = D_\beta \left( \frac{\partial^2 l}{\partial \theta \partial \beta} \right)^{-1} \frac{\partial^2 l}{\partial \beta \partial z^\top}$$

### Resíduos quantílicos normalizados

O resíduo quantílico normalizado ([DUNN; SMYTH, 1996](#)), é definido como:

$$r_{q,i} = \Phi^{-1}\{F(z_i); \hat{\mu}_i, \hat{\phi}\}$$

em que  $\phi$  é a função de distribuição acumulada da normal padrão, e  $F(z_i)$  é a função de distribuição acumulada da distribuição Beta. Se  $F$  é distribuída continuamente,  $F(z_i)$  é uniformemente distribuída no intervalo unitário.

## 2.6 Proposta de metodologia

Os dados multipartidários apresentam diversos desafios para análise e interpretação dos votos. Devido às características de dados composicionais, é necessário realizar transformações, como a transformação de razão logarítmica centrada para eliminar as restrições de soma constante e transferir os dados de um espaço simplex a um espaço Euclidiano.

Nesse contexto, em esta dissertação propomos uma metodologia que consiste no desenvolvimento de métodos estatísticos para facilitar a interpretação desses dados multipartidários.

A metodologia apresenta as seguintes oito etapas:

1. Coleta de dados de votos multipartidários.
2. Transformação dos votos em proporções.
3. Aplicação de transformação em razões logarítmicas das proporções de votos.
4. Aplicação da análise de componentes principais (ACP) para dados composicionais transformados.
5. Identificação de escores de votos usando ACP.
6. Transformação de escores de votos ao intervalo unitário.
7. Identificação de covariáveis ou variáveis explicativas.
8. Aplicação de um modelo de regressão com resposta limitada para os escores transformados.

Na etapa 1, a abordagem consiste em coletar informações dos dados de uma região, estado ou país, nos quais os votos podem variar em quantidade. Na etapa 2, é recomendável transformar esses votos em proporções no intervalo  $(0, 1)$ . Essas proporções indicam uma soma constante de 1 para cada circunscrição, respeitando as restrições dos dados composicionais.

Na etapa 3, aplicamos a transformação de razão logarítmica centrada para eliminar as restrições da soma constante e do espaço simplex restrito. Contudo, os dados apresentam uma estrutura multidimensional que pode complicar a análise. Na etapa 4, para lidar com a multidimensionalidade, utilizamos a Análise de Componentes Principais (ACP) como uma ferramenta para a redução de dimensões. A ACP permite visualizar os dados em um espaço euclidiano, onde as cargas e os escores descrevem as relações entre as variáveis e componentes.

Na etapa 5, identificamos os escores, que refletem a direção dos votos para diferentes partidos políticos, tornando-se uma variável que indica o comportamento das circunscrições em relação à preferência por um partido político. Na etapa 6, dado que os escores apresentam variabilidades diferentes em cada componente, dificultando a interpretabilidade, uma abordagem comum é a normalização, transformando os escores para valores no intervalo  $(0, 1)$ .

Portanto, esses escores normalizados tornam-se o foco do estudo e representam a variável resposta. Na etapa 7, são necessárias variáveis explicativas para o modelo relacionado aos dados, descrevendo a população e auxiliando na escolha por um partido político. Na etapa 8, aplicamos um modelo de regressão com resposta limitada, onde a variável resposta escore está no intervalo  $(0, 1)$ .

Uma ferramenta útil para visualizar a metodologia com suas respectivas etapas no processo e as interações com as etapas anteriores são os diagramas. Na Figura 2, mostramos o diagrama correspondente que propõe as etapas a serem seguidas para analisar e interpretar os dados multipartidários.

A metodologia utiliza a análise de componentes principais como método de redução de dimensões, embora outros métodos, como a análise de fatores, possam ser aplicados. As variáveis explicativas podem variar conforme a região, estado ou circunscrição, identificando quais variáveis exercem maior influência nos votos. Por fim, ao aplicar um modelo de regressão com resposta limitada, diversas funções de ligação podem ser utilizadas, tais como logito, probito, log-log e complementar log-log.

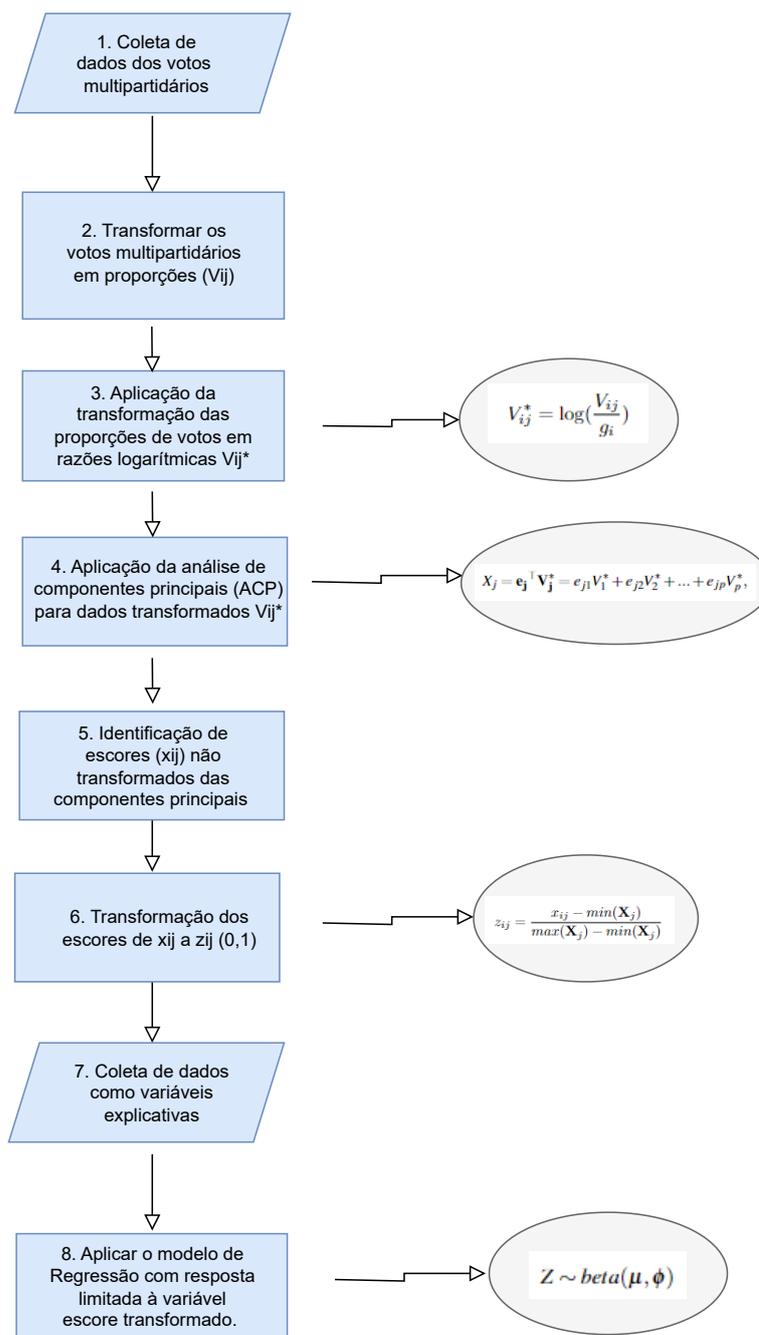


Figura 2 – diagrama geral da metodologia

## 2.7 Revisão de cópulas

O termo "cópulas" foi inicialmente introduzido por Sklar (1959), estabelecendo os fundamentos dessas funções e foram criadas com duas finalidades: procurar uma medida de dependência e fornecer um ponto de partida na construção de distribuições bivariadas. Cópulas são essenciais para relacionar a distribuição multivariada de variáveis com suas distribuições marginais dessas variáveis (NELSEN, 2006), assim como vários estudos foram desenvolvidos com o objetivo de obter distribuições multivariadas mais complexas, onde as distribuições univariadas diferem, e busca-se uma relação entre elas mediante uma estrutura de dependência.

De acordo com [Joe \(2014\)](#), sob o mesmo ponto de vista, as cópulas são distribuições multivariadas com marginais univariadas seguindo uma distribuição  $U(0, 1)$ . Portanto, se  $C$  é uma cópula, então  $C$  é uma distribuição de um vetor de variáveis aleatórias  $U(0, 1)$ .

Simultaneamente, várias famílias de cópulas emergiram para a construção de distribuições multivariadas em diferentes aplicações, sendo as mais notáveis as cópulas elípticas, derivadas das distribuições multivariadas normal e t-student, como a Cópula Gaussiana e Cópula t-Student; e as cópulas arquimedianas, provenientes de uma função geradora que inclui comumente Cópulas de Gumbel, Clayton e Frank.

Dentre os estudos e aplicações de cópulas em diversas áreas, destacam-se os trabalhos mais desenvolvidos em finanças e economia. Por exemplo, [Embrechts, Lindskog e McNeil \(2001\)](#) abordaram modelos de dependência com cópulas na Gestão Integrada de Riscos, enquanto [Bouyé et al. \(2000\)](#) exploraram aplicações em problemas financeiros, como pontuação de crédito, modelagem de retorno de ativos e medida de risco. Além disso, [Denuit, Purcaru e Keilegom \(2006\)](#) utilizaram cópulas para modelar riscos múltiplos em seguros. Contudo, as cópulas também encontraram aplicação em outras áreas, como o modelo multivariado de demanda sísmica de concreto proposto por [Goda e Tesfamariam \(2015\)](#), a aplicação de cópulas gaussianas em Ecologia e Biologia Evolutiva [Prates et al. \(2015\)](#) para a construção de processos Markov Gamma ou Beta, o uso de cópulas arquimedianas em dados de sobrevivência com observações censuradas por [Biondo e Suzuki \(2016\)](#), a exploração da relação entre renda e democracia através de cópulas por [Paleologou \(2023\)](#) e modelos multivariados utilizando cópulas em marketing, conforme desenvolvido por [Danaher e Smith \(2011\)](#).

Em resumo, as cópulas apresentam uma ampla gama de aplicações, oferecendo diversas vantagens, como a obtenção de distribuições multivariadas e a modelagem da estrutura de dependência. Para ilustrar, consideremos a construção de uma distribuição bivariada. Inicialmente contemplamos distribuições marginais acumuladas que são transformadas em distribuições uniformes padrão  $[0, 1]$ . Consequentemente, obtemos uma distribuição bivariada e um parâmetro de dependência, dado que as variáveis possuem as mesmas distribuições marginais uniformes.

#### *Teorema de Sklar*

O teorema proposto por [Sklar \(1959\)](#) estabelece que, para uma função de distribuição multivariada acumulada  $H(x_1, \dots, x_d) = P(X_1 \leq x_1, \dots, X_d \leq x_d)$  de um vetor aleatório  $X_1, \dots, X_d$ , com marginais  $F_{X_i}(x_i) = P(X_i \leq x_i)$ , existe uma função copula d-dimensional  $C$  tal que:

$$H(x_1, \dots, x_d) = C(F_{X_1}(x_1), \dots, F_{X_d}(x_d))$$

Se as marginais  $F_{X_i}(x_i)$  para  $i = 1, \dots, d$  são contínuas, então  $C$  é única; caso contrário,  $C$  é unicamente determinada no conjunto  $\text{ran}F_{X_1} \times \text{ran}F_{X_2} \times \dots \times \text{ran}F_{X_d}$ .

Outra forma de expressar o teorema é considerando  $H$  uma distribuição multivariada

$d$ -dimensional com marginais contínuas  $F_{X_1}(x_1), \dots, F_{X_d}(x_d)$ . A copula  $C$  é determinada para todo  $u_i \in (0, 1)$  como:

$$C(\mathbf{u}) = H(F_{X_1}^{-1}(u_1), \dots, F_{X_d}^{-1}(u_d))$$

onde,  $F_{X_1}^{-1}, \dots, F_{X_d}^{-1}$  são as inversas de  $F_{X_1}, \dots, F_{X_d}$ .

Portanto, o teorema permite obter a distribuição multivariada considerando as distribuições marginais por meio de copulas. Além disso, possibilita modelar a copula adequada para as variáveis e estimar a força de dependência entre elas por meio de estruturas de dependência.

### Propriedades das cópulas

As cópulas apresentam propriedades fundamentais que destacam sua importância:

i)  $C(\mathbf{u}) = u_i$ , quando todas as coordenadas, exceto  $u_i$ , são iguais a 1. Isso é representado por:  
 $C(1, 1, \dots, u_i, \dots, 1, 1) = u_i, \forall i = 1, \dots, d, u_i \in [0, 1]$

ii) Seja  $\mathbf{u} = [u_1, \dots, u_d]$ , onde  $u_i = F_{X_i}(x_i)$ . Se  $u_i = 0$  para  $i \leq n$  (pelo menos uma das coordenadas  $u$  é igual a 0), então

$$C(u_1, \dots, u_d) = 0.$$

iii) Dado que  $C(u_1, \dots, u_d)$  é limitada, temos  $0 \leq C(u_1, \dots, u_d) \leq 1$ . A propriedade representa o limite da distribuição conjunta acumulada no rango  $[0, 1]$ .

iv) Como  $C(u_1, \dots, u_d)$  é  $d$ -crescente, o volume do intervalo  $d$ -dimensional é não negativo,  $\forall (a_1, \dots, a_n)(b_1, \dots, b_d) \in [0, 1]^d$ , onde  $a_i \leq b_i$

$$\sum_{i_1=1}^2 \dots \sum_{i_d=1}^2 (-1)^{i_1 + \dots + i_d} C(x_{1i_1}, x_{2i_2}, \dots, x_{di_d}) \geq 0 \quad (2.8)$$

Com  $x_{j1} = a_j$  e  $x_{j2} = b_j, j = 1, \dots, d$

v) Sejam as variáveis aleatórias  $X_1, \dots, X_d$ , com funções de distribuição marginal  $F_{X_1}(x_1), \dots, F_{X_d}(x_d)$  e a função de distribuição acumulada conjunta  $H(x_1, \dots, x_d)$ . Defina  $u_i = F_{X_i}(x_i)$  para  $i = 1, \dots, d$ . As variáveis  $X_1, \dots, X_d$  são independentes se e somente se  $H(x_1, \dots, x_d) = \prod_{i=1}^d F_{X_i}(x_i)$  (ANJOS *et al.*, 2004). Portanto, a cópula  $C(u_1, \dots, u_d)$  é denominada cópula produto ou independente, sendo definida como:

$$C(u_1, \dots, u_d) = \prod_{i=1}^d u_i$$

Para ilustrar a equação (2.8), consideremos o caso em que  $d = 2$ , onde  $(a_1, a_2), (b_1, b_2) \in [0, 1]^2$  e  $a_1 \leq a_2, b_1 \leq b_2$ . Então, temos que:

$$\sum_{i_1=1}^2 \sum_{i_2=1}^2 (-1)^{i_1 + i_2} C(x_{1i_1}, x_{2i_2}) \geq 0.$$

Portanto,

$$C(a_1, a_2) - C(a_1, b_2) - C(b_1, a_2) + C(b_1, b_2) \geq 0$$

Da mesma forma, quando  $d = 3$ , obtemos o seguinte:

$$C(a_2, b_2, c_2) - C(a_2, b_2, c_1) - C(a_1, b_2, c_2) + C(a_2, b_1, c_1) + C(a_1, b_2, c_1) + C(a_1, b_1, c_2) - C(a_1, b_1, c_1) \geq 0; (a_1, a_2), (b_1, b_2), (c_1, c_2) \in [0, 1]^3$$

### Limite de Fréchet-Hoeffding

De maneira geral, considere que para cada  $C(u_1, \dots, u_d)$  e  $(u_1, \dots, u_d)$  em  $[0, 1]^d$ , o limite é definido por:

$$W(u_1, \dots, u_d) \leq C(u_1, \dots, u_d) \leq M(u_1, \dots, u_d); i \geq 2$$

onde  $W(u_1, \dots, u_d) = \max\{1 - d + \sum_{i=1}^d u_i, 0\}$  é uma cópula quando  $i = 2$  e o limite inferior Fréchet-Hoeffding, e  $M(u_1, \dots, u_d) = \min(u_1, \dots, u_d)$  é uma cópula e o limite superior de Fréchet-Hoeffding.

Para ilustrar, considere uma cópula  $C$  para  $u_1, u_2 \in I$ , seguidamente temos (NELSEN, 2006) que,

$$W(u_1, u_2) = \max(u_1 + u_2 - 1, 0) \leq C(u_1, u_2) \leq M(u_1, u_2) = \min(u_1, u_2),$$

consequentemente, pelo Teorema de Sklar, se  $x$  e  $y$  são variáveis aleatórias com distribuição conjunta  $H$  e marginais  $F$  e  $G$  respectivamente, então,

$$\max(F(x) + G(y) - 1, 0) \leq H(x, y) \leq \min(F(x), G(y))$$

em que  $W$  e  $M$  são copulas, e os limites são funções de distribuição conjunta, denominadas limites de Fréchet-Hoeffding para a função de distribuição conjunta  $H$  com marginais  $W$  e  $M$ .

## Classes de cópulas

### a. Cópulas elípticas

As cópulas elípticas formam uma classe significativa derivada das distribuições elípticas, como a distribuição multivariada normal e t-Student. Algumas das características fundamentais incluem a flexibilidade na dependência positiva e negativa, bem como a dependência nas caudas. No caso das cópulas normais, temos caudas leves, enquanto nas cópulas t-Student, as caudas são pesadas. Essa flexibilidade faz com que as cópulas elípticas sejam aplicáveis em uma variedade de contextos, incluindo finanças, seguros e gestão de riscos. Para obter mais detalhes sobre essas aplicações, consulte Cherubini, Luciano e Vecchiato (2004).

### Cópula Gaussiana

A cópula Gaussiana consiste na combinação conjunta de distribuições marginais univariadas normais. Entre suas características mais relevantes, destacam-se a simetria no espaço e o fator de correlação associado à distribuição multivariada. Além disso, ela apresenta uma baixa dependência nas caudas.

Essa cópula foi implementada em diversas áreas, como finanças, gestão de riscos e riscos de crédito. No entanto, mostrou limitações na estimativa de dependências complexas. Um exemplo disso é mencionado em [Watts \(2016\)](#), que descreve a crise econômica global de 2008, na qual cópulas gaussianas foram utilizadas para modelar risco financeiro. No entanto, o modelo subestimou o risco de cauda e as dependências entre variáveis essenciais.

De acordo com [Cherubini, Luciano e Vecchiato \(2004\)](#), a cópula Gaussiana multivariada é denotada da seguinte maneira:

$$C_{\Sigma}(u) = \phi_{\Sigma}(\phi^{-1}(u_1), \dots, \phi^{-1}(u_d))$$

em que  $\phi^{-1}$  representa a inversa da distribuição normal estandar  $\phi$ , e  $\phi_{\Sigma}$  denota a distribuição normal estandar multivariada com matriz de correlação simétrica e positiva definida  $\Sigma$ . A densidade da cópula Gaussiana ([CZADO, 2019](#)) é dada como:

$$c(\mathbf{u}; \Sigma) = |\Sigma|^{-\frac{1}{2}} \exp \left\{ \frac{1}{2} \mathbf{x}^{\top} (\mathbf{I}_d - \Sigma^{-1}) \mathbf{x} \right\}$$

em que  $\mathbf{x} = (x_1, \dots, x_d)^{\top} \in \mathbb{R}^d$ , e  $x_i = \phi^{-1}(u_i)$  para  $i = 1, \dots, d$ .

### Cópula T-Student

A cópula t-Student consiste na combinação conjunta de distribuições marginais univariadas t-Student. Além de ser considerada uma distribuição elíptica, ela é simétrica e sua caracterização envolve a utilização de uma matriz de correlação e os graus de liberdade. Essa cópula é frequentemente aplicada em áreas como riscos financeiros, seguros e estudos ambientais. De acordo com [Cherubini, Luciano e Vecchiato \(2004\)](#), a cópula t-Student é denotada da seguinte maneira:

$$C_{\Sigma}(u) = t_{\Sigma, \nu}(t_{\Sigma, \nu}^{-1}(u_1), \dots, t_{\Sigma, \nu}^{-1}(u_d))$$

Aqui,  $t^{-1}$  representa a inversa da distribuição t-Student,  $\Sigma$  é uma matriz parâmetro de escala em  $[-1, 1]^{d \times d}$ , e  $\nu$  denota os graus de liberdade. Se  $d = 2$ , A densidade da cópula t-Student bivariada é dada por:

$$c(\mu_1, \mu_2; \nu, \Sigma) = \frac{t(t_{\nu}^{-1}(\mu_1), t_{\nu}^{-1}(\mu_2); \nu, \Sigma)}{t_{\nu}(t_{\nu}^{-1}(\mu_1))t_{\nu}(t_{\nu}^{-1}(\mu_2))}$$

Tabela 1 – Cóopulas bivariadas Elipticas

| Cóopula   | $C(u, v; \theta, \zeta)$                                            | $\theta$                                    | Transformação                                       | Kendall's $\tau$                 |
|-----------|---------------------------------------------------------------------|---------------------------------------------|-----------------------------------------------------|----------------------------------|
| Normal    | $\phi_2(\phi^{-1}(u), \phi^{-1}(v); \theta)$                        | $\theta \in [-1, 1]$                        | $\tanh^{-1}(\theta)$                                | $\frac{2}{\pi}(\arcsin(\theta))$ |
| T-Student | $t_{2, \zeta}(t_{\zeta}^{-1}(u), t_{\zeta}^{-1}(v); \theta, \zeta)$ | $\theta \in [-1, 1], \zeta \in (2, \infty)$ | $\tanh^{-1}(\theta), \log(\zeta - 2 - \varepsilon)$ | $\frac{2}{\pi} \arcsin(\theta)$  |

*b.* Cóopula Arquimedianas

A família de cóopulas arquimedianas desempenha um papel fundamental em várias aplicações devido à facilidade com que podem ser construídas, oferecendo uma ampla gama de famílias e propriedades viáveis (NELSEN, 2006). A obtenção dessas cóopulas é realizada por meio de uma função geradora  $\varphi$  (gerador aditivo) em que seja  $\varphi^{[-1]}$  uma função pseudo-inversa de  $\varphi$  com  $\text{Dom}\varphi^{[-1]} = [0, \infty]$  e  $\text{Ran}\varphi^{[-1]} = \mathbf{I}$ , dada por,

$$\varphi^{[-1]} = \begin{cases} \varphi^{-1}(t), & 0 \leq t \leq \varphi(0) \\ 0, & \varphi(0) \leq t \leq \infty \end{cases}$$

As funções da forma  $C_{\varphi} = \varphi^{[-1]}(\varphi(u) + \varphi(v))$  para  $u, v$  em  $[0, 1]$ , são denominadas cóopulas Arquimedianas e  $\varphi$  é chamada função geradora. No caso em que  $\varphi^{[-1]} = \varphi^{-1}$  e  $C_{\varphi} = \varphi^{-1}(\varphi(u) + \varphi(v))$ , dizemos que é uma cóopula Arquimediana estrita.

Para ilustrar, consideremos a função  $\varphi(t) = -\log t$  para  $t$  em  $[0, 1]$ . Dado que  $\varphi(0) = \infty$ ,  $\varphi$  é estrita. Portanto,  $\varphi^{[-1]}(t) = \varphi^{-1}(t) = \exp(-t)$  e a  $C$  gerada é,

$$C(u, v) = \exp(-[(-\ln u) + (-\ln v)]) = uv = \prod(uv).$$

Conseqüentemente,  $\prod(uv)$  é uma cóopula arquimediana estrita.

A cóopula arquimediana  $C$  com a função geradora  $\varphi$  (NELSEN, 2006) apresenta as seguintes características:

- $C$  possui simetria, portanto  $c(u_1, u_2) = c(u_2, u_1)$  para todo  $u_1, u_2$  em  $\mathbf{I}$ ,
  - $C$  é associativa, ou seja,  $C(C(u_1, u_2), u_3) = C(u_1, C(u_2, u_3))$  para todo  $u_1, u_2, u_3$  em  $\mathbf{I}$ .
- Isso implica que podemos obter a mesma distribuição conjunta associando diferentes pares de variáveis,

- Se  $a > 0$  é uma constante, então  $a\varphi$  é um gerador de  $C$ .

Na seguinte Tabela 2 mostramos os diferentes tipos de cóopula bivariadas classificados como arquimedianas, em que  $D_1(\theta) = \frac{1}{\theta} \int_0^{\theta} \frac{t}{\exp t - 1}$  é a função Debye e  $D_2(\theta) = \int_0^{\theta} t \log(t) (1 - t)^{\frac{2(1-\theta)}{\theta}}$ .

**Dependência**

A distribuição conjunta revela a interação entre as variáveis, no entanto, não oferece informações específicas sobre a natureza da relação entre elas, seja positiva ou negativa. Por

Tabela 2 – Principais famílias de cópulas arquimidianas

| Cópula          | $C(u, v; \theta)$                                                                       | $\theta$                        | Transformação                    | Kendall's $\tau$                                                   |
|-----------------|-----------------------------------------------------------------------------------------|---------------------------------|----------------------------------|--------------------------------------------------------------------|
| Clayton         | $(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}$                                   | $\theta \in (0, \infty)$        | $\log(\theta - \varepsilon)$     | $\frac{\theta}{\theta+2}$                                          |
| Frank           | $-\theta^{-1} \log\{1 + (e^{-\theta u} - 1)\frac{e^{-\theta v} - 1}{e^{-\theta} - 1}\}$ | $\theta \in \mathbb{R} - \{0\}$ | -                                | $1 - \frac{4}{\theta}(1 - D_1(\theta))$                            |
| Gumbel          | $\exp[-\{(-\log u)^\theta + (-\log v)^\theta\}^{\frac{1}{\theta}}]$                     | $\theta \in [1, \infty)$        | $\log(\theta - 1)$               | $1 - \frac{1}{\theta}$                                             |
| Ali-Mikhail-Haq | $\frac{uv}{1 - \theta(1-u)(1-v)}$                                                       | $\theta \in [-1, 1]$            | $\tanh^{-1}(\theta)$             | $-\frac{2}{3\theta^2}\theta + (1 - \theta)^2 \log(1 - \theta) + 1$ |
| Joe             | $1 - \{(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\}^{\frac{1}{\theta}}$     | $\theta \in (1, \infty)$        | $\log(\theta - 1 - \varepsilon)$ | $1 + \frac{4}{\theta^2}D_2(\theta)$                                |

consiguiente, foram estudadas diferentes medidas de dependência na literatura, como medidas de quantificação, incluindo o coeficiente de correlação linear, que avalia o grau de correlação linear; o coeficiente de Kendall ( $\tau$ ) e o coeficiente de Spearman que medem o grau de concordância (ANJOS *et al.*, 2004) com valores variando entre -1 e 1. Um valor de -1 indica uma dependência perfeitamente negativa, enquanto um valor de 1 indica uma dependência perfeitamente positiva. Valores entre esses extremos indicam diferentes formas de associação entre as variáveis.

### Correlação de Pearson

A correlação mede se os conjuntos de variáveis estão linearmente relacionados e é definida como a razão entre a covariância e os desvios padrões. A fórmula para o coeficiente de correlação ( $r$ ) é a seguinte:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Aqui,  $X_i$  e  $Y_i$  representam as variáveis, onde  $i = 1, \dots, n$ , e  $\bar{X}$  e  $\bar{Y}$  denotam as médias de cada variável, respectivamente. Os valores de  $r$  variam entre -1 e 1, onde valores próximos de -1 indicam uma correlação negativa, valores próximos de 1 indicam uma correlação positiva, e o valor 0 indica que as variáveis são independentes.

### Concordância

O termo significa que um par de variáveis é considerado concordante quando uma grande quantidade de valores de uma variável está associada a uma grande quantidade de valores da outra variável, e pequenos valores de uma variável estão relacionados a pequenos valores da outra variável. Portanto, ao considerarmos duas observações,  $(x_i, y_i)$  e  $(x_j, y_j)$ , de um vetor de variáveis contínuas  $(X, Y)$ , dizemos que as observações são concordantes se  $x_i < x_j$  e  $y_i < y_j$  ou se  $x_i > x_j$  e  $y_i > y_j$ . Da mesma forma, as observações são discordantes se  $x_i < x_j$  e  $y_i > y_j$ , ou se  $x_i > x_j$  e  $y_i < y_j$ . Uma forma alternativa é:  $(x_i, y_i)$  e  $(x_j, y_j)$  são considerados concordantes se  $(x_i - x_j)(y_i - y_j) > 0$  e discordantes se  $(x_i - x_j)(y_i - y_j) < 0$ .

### Coefficiente de Kendall

Seja  $(X, Y)$  um vetor aleatório com cópula  $C$ . O coeficiente de Kendall  $\tau$  para  $X$  e  $Y$ , em termos de cópula, é definido como:

$$\tau(X, Y) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1$$

Além disso, a integral pode ser interpretada como o valor esperado da função  $C(U, V)$  de variáveis aleatórias uniformes  $(0, 1)$ ,  $U$  e  $V$ , com função de distribuição conjunta  $C$ :

$$\tau_C = 4E(C(U, V)) - 1$$

### *Coefficiente de Spearman*

Seja um vetor  $(X, Y)$  um vetor aleatório com cópula associada  $C$ . O coeficiente de Spearman é denotado da seguinte forma (NELSEN, 2006):

$$\rho_{X, Y} = 3Q(C, \pi) = 12 \int_0^1 \int_0^1 uv dC(u, v) - 3 = 12 \int_0^1 \int_0^1 C(u, v) dudv - 3$$

, onde o coeficiente 3 é uma constante de normalização, considerando que  $Q(C, \pi) \in [-\frac{1}{3}, \frac{1}{3}]$ .

## ***Criterios de Seleção de modelos***

### *Criterio de informação Akaike (AIC)*

Baseado no conceito de informação de entropia, escolhamos o modelo com um valor de AIC menor. O critério mitiga o risco de sobreajuste (overfitting) introduzindo um termo de penalidade  $2d$  que aumenta com o número de parâmetros, permitindo descartar modelos desnecessários. O critério AIC é dado pela fórmula (AKAIKE, 1974):

$$AIC = -2\log(\hat{L}) + 2d$$

onde  $\hat{L}$  é o valor de maximização da função de verossimilhança e  $2d$  é a penalidade pelo número de parâmetros.

### *Criterio de informação Bayesiana (BIC)*

O BIC (SCHWARZ, 1978) está relacionado ao critério AIC no qual o modelo com o menor valor é considerado o melhor. Este critério, é fundamentado na função de verossimilhança. A fórmula do critério BIC é a seguinte:

$$BIC = k \log(n) - 2\log(\hat{L})$$

em que  $\hat{L}$  representa o valor de maximização da função de verossimilhança do modelo.  $n$  é o número de observações e  $k$  é o número de parâmetros a serem estimados.

---

# UMA ABORDAGEM ESTATÍSTICA PARA A ANÁLISE DAS ELEIÇÕES PRESIDENCIAIS DO PERU 2021

---

---

## 3.1 Introdução

Os dados utilizados nesta aplicação consistem nos resultados (número de votos) do primeiro turno das eleições presidenciais peruanas de 2021 para as 24 regiões e a província constitucional do país, Callao. As eleições peruanas estavam marcadas para 11 de abril de 2021 e determinaram a escolha do presidente, do vice-presidente e a composição do parlamento, que conta com 130 membros. Em caso de nenhum candidato obter mais da metade dos votos no primeiro turno, ocorre um segundo turno entre os dois primeiros candidatos com o maior número de votos.

O processo eleitoral contou com a participação de dezoito organizações eleitorais, a maior quantidade desde as eleições peruanas de 2006. Alguns dos candidatos notáveis foram Pedro Castillo do Peru Libre, Keiko Fujimori do Fuerza Popular, Rafael López Aliaga do Renovacion Popular, Hernando de Soto do Avanza país, Yonhy Lescano do Accion Popular, entre outros.

Os partidos políticos participantes possuem diferentes orientações ideológicas. Peru Libre e Accion Popular adotam uma postura socialista com um espectro político progressista, Fuerza Popular possui uma abordagem neoliberal com espectro conservador, Renovacion Popular adota uma postura liberal também com espectro conservador, e Avanza país é um partido neoliberal com espectro conservador.

Os resultados do primeiro turno de acordo com a [ONPE \(2021\)](#) foram os seguintes: Perú libre (PL) com 18,92% dos votos, Fuerza Popular (FP) com 13,4%, Renovación Popular(RP) (11,75%), Avanza país(AVP) (11,62%), Accion Popular(ACCP) (9,071%), Juntos por el Perú

(7,86%), Alianza por el progreso (6,021%), Victoria Nacional (5,65%) e outros (15,68%). Esses resultados evidenciam, mais uma vez, a alta popularidade do partido Fuerza Popular, que conseguiu avançar para o segundo turno com uma expressiva porcentagem de votos. No entanto, questões políticas como corrupção e acusações de violação dos direitos humanos relacionadas ao governo de Alberto Fujimori, pai da candidata Keiko Fujimori, geraram uma atitude antifujimorista em diversas regiões do Peru. No segundo turno, realizado em 19 de julho de 2021, Pedro Castillo, do partido PL foi escolhido como presidente, com aproximadamente 50,12% dos votos.

Os dados do primeiro turno das eleições, apresentados na Tabela 3, incluem os primeiros cinco partidos denotados como PL, FP, RP, AVP, AP, além de uma categoria adicional OUTROS, que engloba outras organizações com pouca porcentagem de votos.

Tabela 3 – Dados das eleições peruanas de 2021

|    | Região        | Código | PL      | FP      | RP      | AVP     | ACCP    | OUTROS    | Votos Válidos |
|----|---------------|--------|---------|---------|---------|---------|---------|-----------|---------------|
| 1  | Amazonas      | AMA    | 34.464  | 17.815  | 8.274   | 4.433   | 12.703  | 54.510    | 132.199       |
| 2  | Áncash        | ANC    | 110.620 | 67.394  | 42.312  | 34.562  | 38.911  | 177.986   | 471.785       |
| 3  | Apurímac      | APU    | 88.812  | 10.879  | 7.768   | 6.531   | 15.649  | 36.547    | 166.186       |
| 4  | Arequipa      | ARE    | 256.224 | 40.216  | 71.053  | 148.793 | 88.708  | 190.717   | 795.711       |
| 5  | Ayacucho      | AYA    | 130.224 | 17.751  | 11.490  | 8.995   | 20.315  | 61.775    | 250.550       |
| 6  | Cajamarca     | CAJ    | 232.418 | 54.962  | 31.129  | 25.156  | 38.677  | 135.120   | 517.462       |
| 7  | Callao        | CAL    | 33.750  | 79.699  | 78.066  | 78.920  | 34.965  | 219.867   | 525.267       |
| 8  | Cusco         | CUS    | 232.178 | 27.132  | 29.618  | 40.423  | 60.659  | 218.023   | 608.033       |
| 9  | Huancavelica  | HUV    | 79.895  | 8.449   | 5.060   | 4.591   | 16.727  | 32.665    | 147.387       |
| 10 | Huánuco       | HUC    | 110.978 | 32.827  | 33.787  | 15.822  | 22.565  | 79.244    | 295.223       |
| 11 | Ica           | ICA    | 56.627  | 62.102  | 46.116  | 39.949  | 39.475  | 161.545   | 405.814       |
| 12 | Junín         | JUN    | 131.438 | 80.057  | 52.599  | 54.124  | 66.214  | 189.666   | 574.098       |
| 13 | La Libertad   | LAL    | 90.324  | 131.866 | 95.973  | 84.566  | 47.322  | 334.797   | 784.848       |
| 14 | Lambayeque    | LAM    | 73.279  | 121.263 | 86.126  | 50.087  | 51.467  | 184.346   | 566.568       |
| 15 | Lima          | LIM    | 416.743 | 754.216 | 870.416 | 871.000 | 362.881 | 2.035.250 | 5.310.506     |
| 16 | Loreto        | LOR    | 15.662  | 52.344  | 16.449  | 18.846  | 34.998  | 175.809   | 314.108       |
| 17 | Madre De Dios | MDD    | 23.945  | 7.278   | 4.041   | 3.996   | 6.601   | 18.713    | 64.574        |
| 18 | Moquegua      | MOQ    | 33.665  | 4.617   | 6.832   | 10.183  | 15.412  | 27.217    | 97.926        |
| 19 | Pasco         | PAS    | 34.187  | 12.607  | 8.009   | 5.102   | 11.871  | 28.220    | 99.996        |
| 20 | Piura         | PIU    | 71.028  | 173.933 | 68.356  | 63.866  | 51.250  | 272.512   | 700.945       |
| 21 | Puno          | PUN    | 292.218 | 17.514  | 15.918  | 21.665  | 175.712 | 92.494    | 615.521       |
| 22 | San Martín    | SAM    | 67.000  | 46.699  | 26.561  | 21.825  | 31.498  | 119.887   | 313.470       |
| 23 | Tacna         | TAC    | 64.521  | 9.363   | 17.842  | 21.000  | 28.696  | 52.847    | 194.269       |
| 24 | Tumbes        | TUM    | 7.613   | 36.403  | 8.799   | 7.123   | 7.046   | 31.257    | 98.241        |
| 25 | Ucayali       | UCA    | 26.339  | 40.510  | 14.981  | 11.124  | 14.359  | 81.057    | 188.370       |

PL: Peru Libre, FP: Fuerza Popular, RP: Renovacion Popular, AVP: Avanza país, AP: Accion Popular e OUTROS

Os dados revelam que a região de Lima (capital do Peru) apresenta a maior quantidade de votos válidos, totalizando 5.310.506 votos, indicando sua significativa influência nas eleições e seu papel crucial na escolha dos governantes. Por outro lado, a região de Madre De Dios registra a menor quantidade de votos válidos, com 64.574, sugerindo uma menor influência nessa região.

Ao analisar os partidos que receberam o maior número de votos, observamos que em regiões como Arequipa, Lima, Puno e Cuzco, localizadas no centro do país, o partido Peru Libre recebeu uma quantidade expressiva de votos. Enquanto isso, em regiões como Lima, La

Libertad, Piura e Lambayeque, localizadas na costa do país, o partido Fuerza Popular recebeu a maior quantidade de votos. Esses resultados indicam padrões regionais distintos nas preferências eleitorais, evidenciando a diversidade de influências e comportamentos de voto em diferentes partes do país.

## 3.2 Metodologia

Dado que buscamos interpretar os votos das regiões do Peru, aplicamos a metodologia proposta nas eleições presidenciais do ano de 2021 considerando as seguintes etapas.

1. Coleta de dados das eleições presidenciais de Peru do ano de 2021.
2. Transformação dos votos do Peru em proporções.
3. Aplicação da transformação em razões logarítmicas das proporções de votos do Peru.
4. Aplicação da análise de componentes principais (ACP) para dados composicionais transformados do Peru.
5. Identificação de escores das regiões de Peru por meio da ACP.
6. Transformação dos escores para o intervalo unitário.
7. Coleta de indicadores de desenvolvimento humano (IDH) do Peru.
8. Aplicação de um modelo de regressão Beta à variável escore transformada.

Na etapa 1, realizamos a coleta de dados dos votos no Peru, evidenciando variabilidades distintas no número de votos por região. Na etapa 2, procedemos à transformação dos dados em proporções  $V_{ij}$ , em que as somas proporcionais dos votos por região totalizam 1. Dada a natureza composicional dos dados proporcionais de votos no Peru, na etapa 3 aplicamos a transformação de razões logarítmicas para obter  $V_{ij}^*$  e remover as restrições de soma constante e espaço restrito simplex.

Em seguida, Na etapa 4, conduzimos a análise de componentes principais para extrair as principais componentes explicativas da variância nos dados. Conseqüentemente, obtemos as cargas dos partidos eleitorais em cada componente e os escores de cada região nas componentes. Na etapa 5, extraímos os escores  $x_{ij}$  em que  $i$  indica as regiões de Peru e  $j$  a componente associada. Na etapa 6, considerando que os escores apresentam variabilidades distintas nas componentes, realizamos a transformação para  $z_{ij}$ , cujos valores variam entre  $(0, 1)$ . O escore transformado é, então, considerado a variável resposta, e o objetivo é interpretar esses valores para entender por que as regiões apresentam esse valor de escore em cada componente, com base em um conjunto de covariáveis que explicam o comportamento político da população.

Dessa forma, na etapa 7 incorporamos os indicadores de desenvolvimento humano do Peru como variáveis explicativas, classificadas em saúde, educação e renda, que são fundamentais para compreender como influenciam no comportamento político. Na etapa 8, aplicamos o modelo de regressão Beta, dada a capacidade para lidar com valores no intervalo (0, 1), considerando os escores como variáveis de resposta e os indicadores de desenvolvimento humano do Peru como covariáveis. A Figura 3 apresenta o diagrama com detalhes sobre o processo.

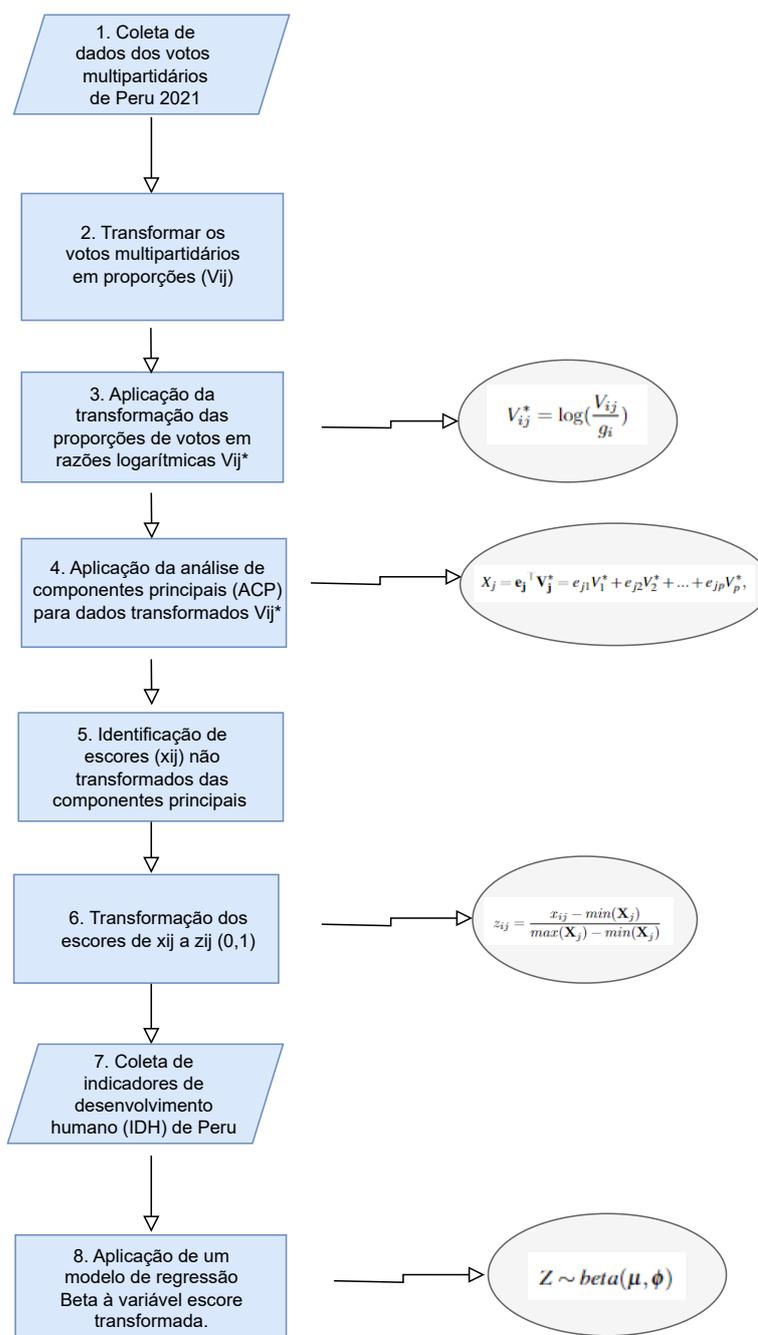


Figura 3 – Diagrama da metodologia aplicada nas eleições peruanas de 2021

Os códigos da metodologia proposta foram desenvolvidos no software R e estão disponíveis na [Código-fonte 1](#). Além disso, podem ser acessado pelo seguinte link: [Bounded regression](#)

model Peru 2021.

### 3.3 Resultados

#### 3.3.1 Matrizes de correlação das eleições Peruanas de 2021

Considerando as três bases de dados obtidas das eleições presidenciais do Peru em 2021, analisamos as correlações das variáveis (organizações políticas participantes) em cada uma delas. Na [Tabela 4](#), são apresentadas as correlações de acordo com o número de votos, onde todas as correlações possuem valores positivos, significando que um aumento em qualquer dos partidos, gera um aumento no outro partido, e como sabemos isso não acontece nos dados multipartidários. Além disso, é importante notar que a correlação é espúria, uma vez que as variáveis são dependentes. Na [Tabela 5](#), são exibidas as correlações das proporções de votos ou dados composicionais. Neste caso, observamos uma tendência diferente, com correlações positivas e negativas. Por exemplo, os partidos PL e FP mostram uma correlação negativa de -0,75, e o partido RP e PL mostram uma correlação negativa de -0,70. Apesar disso, as variáveis continuam sendo dependentes e apresentam a restrição de soma constante um, o que pode levar a resultados inapropriados. Finalmente, na [Tabela 6](#), apresentamos a matriz de correlação considerando a transformação de razão logarítmica das votações proporcionais. Nesse caso, os dados estão no espaço Euclidiano, e a restrição de soma constante de um foi removida. Como resultado, as correlações dos dados com a transformação clr nos conduzem a resultados mais apropriados, possibilitando uma análise adequada do que está ocorrendo com os dados. Consequentemente, a correlação de PL e FP com valor -0,72, ou RP e PL com valor -0,73, tem mais significância no análise dos dados.

Tabela 4 – Matriz de correlação considerando o número de votos no Peru

|        | PL    | FP    | RP    | AVP   | ACCP  | OUTROS |
|--------|-------|-------|-------|-------|-------|--------|
| PL     | 1,000 | 0,595 | 0,646 | 0,679 | 0,837 | 0,659  |
| FP     | 0,595 | 1,000 | 0,984 | 0,970 | 0,868 | 0,987  |
| RP     | 0,646 | 0,984 | 1,000 | 0,994 | 0,893 | 0,994  |
| AVP    | 0,679 | 0,970 | 0,994 | 1,000 | 0,904 | 0,989  |
| ACCP   | 0,837 | 0,868 | 0,893 | 0,904 | 1,000 | 0,899  |
| OUTROS | 0,659 | 0,987 | 0,994 | 0,989 | 0,899 | 1,000  |

#### 3.3.2 Identificando as principais componentes das eleições Peruanas

As principais componentes são aquelas que contêm a maior parte da variância dos dados e explicam as características das eleições peruanas. O objetivo deste estudo é identificá-las. Inicialmente, utilizamos o pacote R para análise composicional conforme proposto por [Boogaart e Tolosana-Delgado \(2008\)](#). Os resultados revelaram duas métricas essenciais: cargas e escores, além da variância acumulada por componente. Normalmente, as duas primeiras componentes

Tabela 5 – Matriz de correlação considerando as proporções de votos no Peru

|        | PL     | FP     | RP     | AVP    | ACCP   | OUTROS |
|--------|--------|--------|--------|--------|--------|--------|
| PL     | 1,000  | -0,749 | -0,699 | -0,510 | 0,400  | -0,816 |
| FP     | -0,749 | 1,000  | 0,438  | 0,062  | -0,506 | 0,519  |
| RP     | -0,699 | 0,438  | 1,000  | 0,698  | -0,480 | 0,380  |
| AVP    | -0,510 | 0,062  | 0,698  | 1,000  | -0,194 | 0,208  |
| ACCP   | 0,400  | -0,506 | -0,480 | -0,194 | 1,000  | -0,494 |
| OUTROS | -0,816 | 0,519  | 0,380  | 0,208  | -0,494 | 1,000  |

Tabela 6 – Matriz de correlação considerando as transformações clr dos votos no Peru

|        | PL     | FP     | RP     | AVP    | ACCP   | OUTROS |
|--------|--------|--------|--------|--------|--------|--------|
| PL     | 1,000  | -0,716 | -0,734 | -0,568 | 0,592  | -0,600 |
| FP     | -0,716 | 1,000  | 0,476  | -0,064 | -0,673 | 0,560  |
| RP     | -0,734 | 0,476  | 1,000  | 0,624  | -0,792 | 0,163  |
| AVP    | -0,568 | -0,064 | 0,624  | 1,000  | -0,330 | 0,012  |
| ACCP   | 0,592  | -0,673 | -0,792 | -0,330 | 1,000  | -0,366 |
| OUTROS | -0,600 | 0,560  | 0,163  | 0,012  | -0,366 | 1,000  |

mostram a maior parte dessa variância, e uma ferramenta útil para visualizar os dados nessas componentes é o biplot (GABRIEL, 1971). O biplot permite observar as cargas das variáveis e os escores das observações nas componentes, onde as cargas das variáveis indicam a força da relação entre as componentes e as variáveis, podendo ser positivas ou negativas, enquanto os escores representam as posições das observações nos fatores.

#### *Análise de principais componentes (ACP) dos dados*

A Tabela 7 mostra as componentes para os dados reais, onde podemos observar cargas positivas e negativas, enquanto espaços nulos indicam cargas próximas de zero. A variância acumulada para a primeira componente é de 88,8%. No entanto, a interpretação não é precisa devido à interdependência das variáveis.

Tabela 7 – Cargas da análise de componentes principais dos dados

| Partidos        | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|-----------------|--------|--------|--------|--------|--------|--------|
| PL              | 0,332  | 0,839  | 0,411  | 0,122  |        |        |
| FP              | 0,418  | -0,312 |        | 0,728  | 0,391  | -0,206 |
| RP              | 0,425  | -0,232 | 0,140  | -0,247 | 0,220  | 0,798  |
| AVP             | 0,427  | -0,168 | 0,186  | -0,617 | 0,236  | -0,565 |
| ACCP            | 0,413  | 0,270  | -0,868 |        |        |        |
| OUTROS          | 0,426  | -0,210 | 0,142  | 0,105  | -0,861 |        |
| Desvio padrão   | 2,309  | 0,754  | 0,255  | 0,160  | 0,075  | 0,055  |
| Var. Proporção  | 0,888  | 0,095  | 0,010  | 0,004  | 0,000  | 0,001  |
| Var. cumulativa | 0,888  | 0,983  | 0,994  | 0,998  | 0,999  | 1,000  |

O biplot na Figura 4 mostra as cargas e os escores dos dados reais nas duas principais componentes, considerando, ademais, que os nomes das regiões são identificados pelas siglas

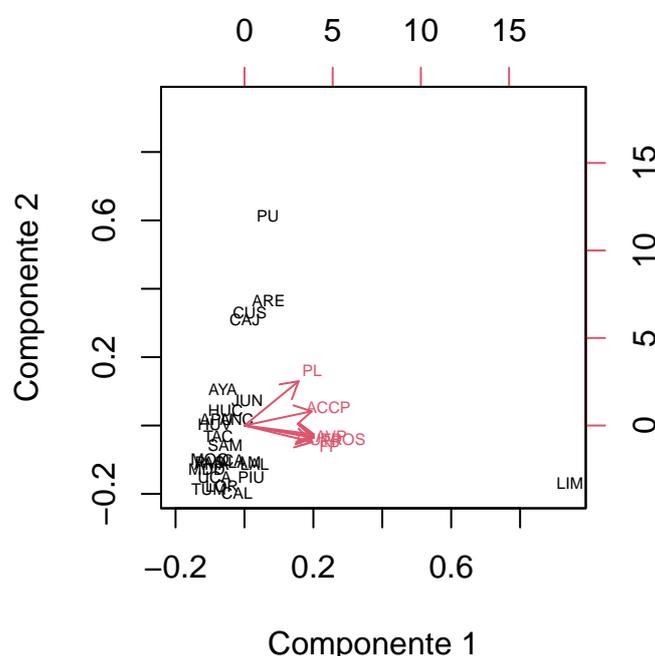


Figura 4 – Biplot da ACP dos dados (98,3% da variância obtida pelas duas primeiras componentes)

codigo estabelecidas por ISO3166 (1998). As flechas representam as cargas, e suas direções estão associadas à contribuição dos partidos na primeira e segunda componente. No entanto, a interpretação não é intuitiva devido às cargas de algumas variáveis se dirigirem em direções semelhantes e às regiões se sobreporem umas às outras.

#### *Análise de componentes principais (ACP) das proporções dos dados*

A Tabela 8 segue a mesma abordagem utilizada na análise das componentes dos dados reais. Neste caso, os dados são proporcionais ou composicionais. As cargas podem ser positivas, negativas ou próximas de zero, indicadas por espaços vazios na tabela.

Tabela 8 – Cargas da análise de componentes principais das proporções dos dados

| Partidos        | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|-----------------|--------|--------|--------|--------|--------|--------|
| PL              | 0,506  |        | 0,412  |        |        | 0,756  |
| FP              | -0,401 | -0,419 |        | -0,656 | -0,309 | 0,366  |
| RP              | -0,436 | 0,400  | 0,189  | -0,250 | 0,723  | 0,172  |
| AVP             | -0,302 | 0,720  |        | 0,109  | -0,582 | 0,195  |
| ACCP            | 0,358  | 0,231  | -0,825 | -0,256 | 0,164  | 0,214  |
| OUTROS          | -0,417 | -0,301 | -0,332 | 0,653  | 0,126  | 0,426  |
| Desvio padrão   | 1,859  | 1,087  | 0,819  | 0,707  | 0,437  | 0,000  |
| Var. Proporção  | 0,576  | 0,197  | 0,112  | 0,083  | 0,032  | 0,000  |
| Var. cumulativa | 0,576  | 0,773  | 0,885  | 0,968  | 1,000  | 1,000  |

O biplot na Figura 5 mostra os dois primeiros componentes de dados proporcionais. A

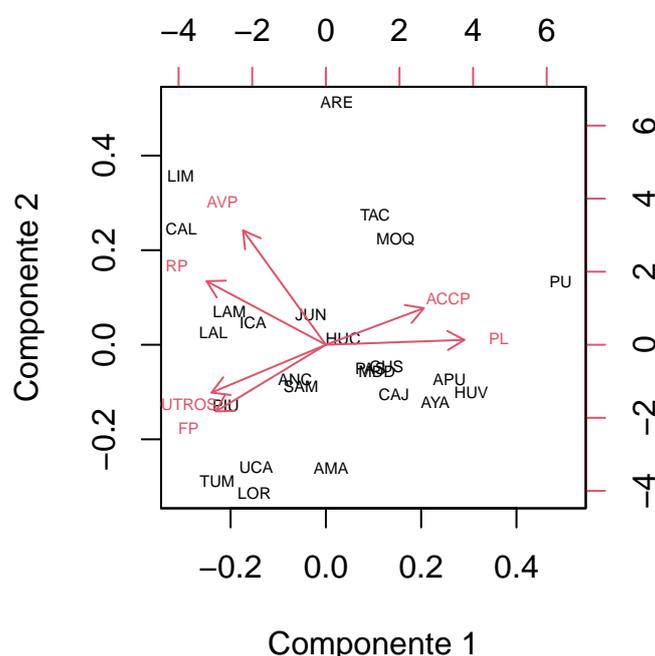


Figura 5 – Biplot da ACP das proporções dos dados (77,3% da variância obtida pelas duas primeiras componentes)

proporção acumulada da variância é 77,3%, valor menor do que a variância acumulada das duas primeiras componentes na análise direta dos dados. Além disso, os partidos políticos foram divididos em três grupos com base em suas posições no biplot: PL e ACCP estão no lado direito, RP e AVP estão no lado superior esquerdo, enquanto FP e OUTROS estão no lado inferior esquerdo.

#### *Análise de componentes principais da transformação clr dos dados*

A Tabela 9 mostra as componentes para os dados transformados. Denotamos a primeira e segunda componente principal com  $x_1$  e  $x_2$ , respectivamente. As cargas exibem valores negativos, positivos ou próximos de zero, indicadas por espaços vazios. Além disso, observamos que as duas primeiras componentes principais explicam 81,1% do total da variância, enquanto as três primeiras explicam 92,8%. Os partidos políticos como PL e ACCP estão positivamente correlacionados com  $x_1$ , enquanto FP e OUTROS estão positivamente correlacionados com  $x_2$ , e RP e AVP estão negativamente correlacionados com  $x_2$ .

O biplot da Figura 6 exhibe as duas primeiras componentes da transformação clr dos dados. É possível observar que as cargas dos partidos FP e OUTROS apontam em direção oposta às de RP e AVP no eixo "y". FP e OUTROS (Partido Nacionalista ou Partido Democrático Somos Peru) representam partidos considerados tradicionais, dado que participaram em diferentes eleições no Peru ao longo de muitos anos. Portanto, propomos que a segunda componente principal  $x_2$  esteja associada à preferência por partidos tradicionais, onde a parte inferior do eixo está relacionada a

Tabela 9 – Cargas da análise de componentes principais da transformação clr dos dados

| Partidos        | Comp.1( $x_1$ ) | Comp.2( $x_2$ ) | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|-----------------|-----------------|-----------------|--------|--------|--------|--------|
| PL              | 0,495           |                 | -0,325 | -0,458 |        | 0,658  |
| FP              | -0,412          | 0,445           | -0,272 | 0,479  | -0,330 | 0,469  |
| RP              | -0,456          | -0,338          | -0,276 |        | 0,723  | 0,282  |
| AVP             | -0,277          | -0,657          | 0,412  |        | -0,456 | 0,336  |
| ACCP            | 0,454           |                 | 0,480  | 0,573  | 0,351  | 0,335  |
| OUTROS          | -0,308          | 0,507           | 0,587  | -0,48  | 0,178  | 0,205  |
| Desvio padrão   | 1,872           | 1,166           | 0,840  | 0,558  | 0,347  | 0,000  |
| Var. Proporção  | 0,584           | 0,226           | 0,117  | 0,051  | 0,020  | 0,000  |
| Var. cumulativa | 0,583           | 0,810           | 0,928  | 0,979  | 1,000  | 1,000  |

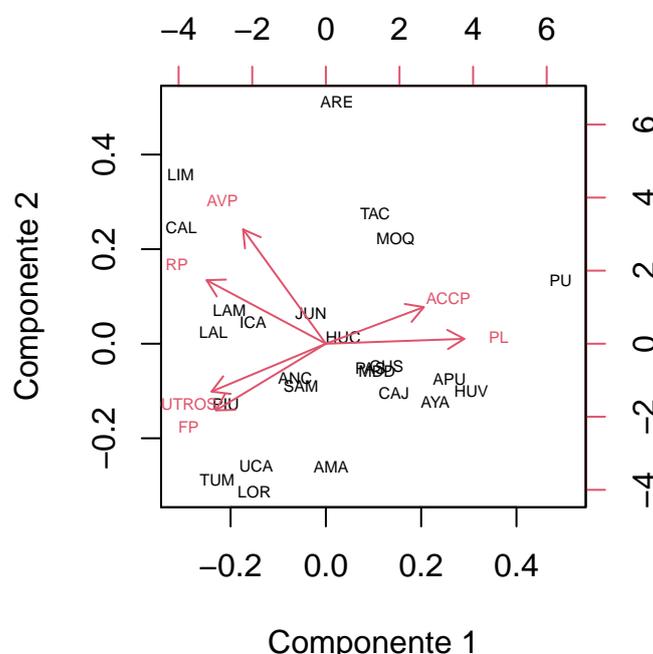


Figura 6 – Biplot da ACP da transformação clr dos dados (81,06% da variância obtida pelas duas primeiras componentes)

partidos menos tradicionais e a parte superior do eixo está associada a partidos tradicionais. Por outro lado, considerando o eixo "x" no biplot, observamos que o partido PL está localizado no lado direito, enquanto RP está no lado esquerdo. Durante o processo eleitoral, esses dois partidos foram opostos em vários aspectos, como economia, visão social, política, entre outros. Portanto, PL representa a política progressista, enquanto RP representa a política conservadora na política peruana. Nesse contexto, propomos que a primeira componente principal  $x_1$  esteja associada à preferência por partidos progressistas.

Por exemplo, é possível observar que as regiões de Piura, Loreto e Ucayali demonstram preferência por partidos tradicionais e conservadores. Por outro lado, a região de Puno opta

por partidos políticos progressistas, enquanto Lima escolheu partidos não tradicionais e conservadores. As regiões de Arequipa e Tacna mostram uma tendência por partidos políticos não tradicionais.

### 3.3.3 Explicando as componentes principais $x_1$ e $x_2$

Em síntese, temos duas componentes principais: a componente  $x_1$  representa a preferência por partidos progressistas, sendo que seu eixo  $x$  indica se a região tem baixa preferência (lado esquerdo) ou alta preferência (lado direito). Por outro lado, a componente  $x_2$  representa a preferência por partidos tradicionais, com seu eixo  $y$  indicando se uma região tem baixa preferência (lado inferior) ou alta preferência (lado superior). Além disso, para avaliar a preferência da região, é necessário focar nos escores dado que implica as posições das regiões e conseqüentemente um aumento ou diminuição dos escores indica a intensidade de preferência.

### 3.3.4 Transformando os escores das componentes principais $x_1$ e $x_2$

A [Figura 6](#) exibe a posição ou escore  $x_{ij}$  para as regiões de Peru, onde  $i$  e  $j$  indicam região e as componentes principais, respectivamente. As coordenadas dos escores em cada componente variam em escalas diferentes e devido à variabilidade, realizamos uma transformação monotônica nos dois escores. finalmente, obtemos valores normalizados entre 0 e 1 por meio da seguinte transformação:

$$z_{ij} = \frac{x_{ij} - \min(\mathbf{x}_j)}{\max(\mathbf{x}_j) - \min(\mathbf{x}_j)}$$

em que,  $0 \leq z_{ij} \leq 1$  denota os escores transformados, e  $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$  é o vetor de escores não transformados para as componentes  $j = 1, 2$  e regiões  $i = 1, \dots, 25$ . A [Tabela 10](#) mostra os valores dos escores transformados para componentes  $x_1$  e  $x_2$  no Peru.

Os valores dos escores não transformados exibem uma variação distinta para cada componente. Em  $x_1$ , Lima apresenta o menor valor, com  $-3,66$ , enquanto Puno exibe o maior valor, atingindo  $4,91$ . Por outro lado, em  $x_2$ , Tacna apresenta um valor de  $-2,35$ , enquanto Amazonas mostra um valor de  $2,12$ . Como resultado, os dois componentes têm variabilidades distintas nos escores, e a transformação pode facilitar a interpretabilidade. Seguidamente, a [Figura 7](#) mostra as densidades dos escores transformados de cada componente principal, com valores variando entre 0 e 1.

### 3.3.5 Proposta do modelo de regressão Beta

Uma vez que os escores transformados variam de 0 a 1, propomos um modelo de regressão no qual tratamos o escore como uma variável resposta limitada dado que tem valores entre 0 e 1. A [Tabela 11](#) apresenta os principais indicadores de desenvolvimento humano do Peru,

Tabela 10 – Escores transformados das componentes  $x_1$  e  $x_2$

| Região        | $x_1$  | $x_2$  | $z_1$ | $z_2$ |
|---------------|--------|--------|-------|-------|
| Amazonas      | 0,287  | 2,116  | 0,460 | 1,000 |
| Áncash        | -0,759 | 0,305  | 0,338 | 0,594 |
| Apurímac      | 2,885  | 0,805  | 0,763 | 0,706 |
| Arequipa      | 1,215  | -2,314 | 0,569 | 0,007 |
| Ayacucho      | 2,177  | 0,201  | 0,681 | 0,571 |
| Cajamarca     | 1,819  | 0,561  | 0,639 | 0,651 |
| Callao        | -3,008 | -1,081 | 0,076 | 0,284 |
| Cusco         | 2,050  | -0,676 | 0,666 | 0,374 |
| Huancavelica  | 3,612  | 1,427  | 0,848 | 0,846 |
| Huánuco       | 0,062  | -0,565 | 0,434 | 0,399 |
| Ica           | -1,534 | 0,118  | 0,248 | 0,552 |
| Junín         | -0,291 | -0,369 | 0,393 | 0,443 |
| La Libertad   | -2,337 | -0,233 | 0,154 | 0,474 |
| Lambayeque    | -2,199 | 0,031  | 0,170 | 0,533 |
| Lima          | -3,660 | -2,269 | 0,000 | 0,017 |
| Loreto        | -2,184 | 2,014  | 0,172 | 0,977 |
| Madre De Dios | 0,946  | 0,134  | 0,537 | 0,556 |
| Moquegua      | 1,729  | -1,292 | 0,629 | 0,236 |
| Pasco         | 0,860  | 0,339  | 0,527 | 0,602 |
| Piura         | -2,644 | 0,608  | 0,118 | 0,662 |
| Puno          | 4,913  | 0,115  | 1,000 | 0,552 |
| San Martín    | -0,452 | 0,453  | 0,374 | 0,627 |
| Tacna         | 2,314  | -2,347 | 0,697 | 0,000 |
| Tumbes        | -3,202 | 0,526  | 0,053 | 0,644 |
| Ucayali       | -1,287 | 1,394  | 0,277 | 0,838 |

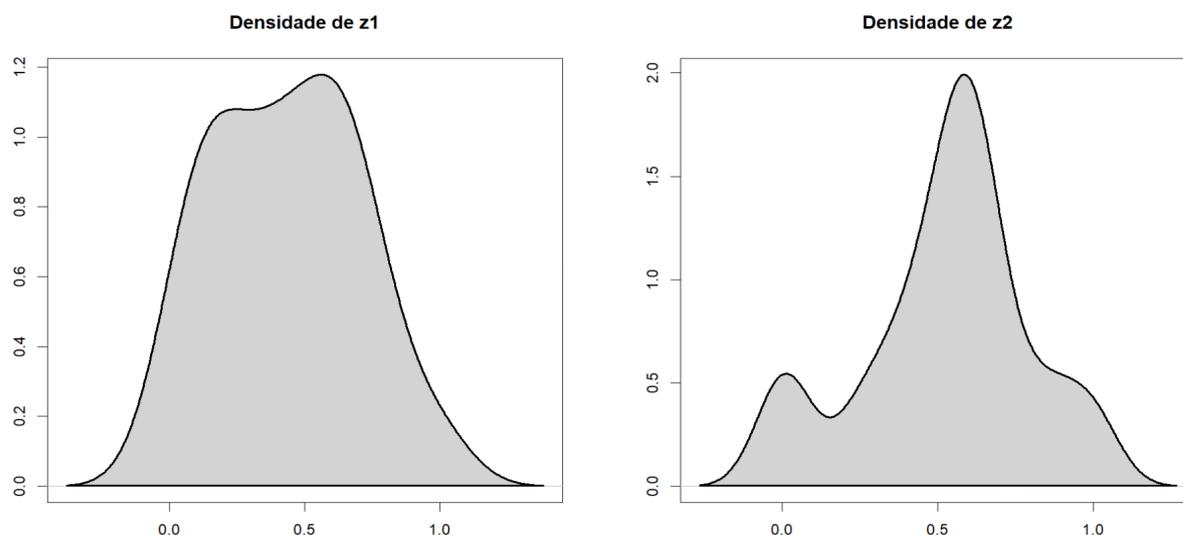


Figura 7 – Densidades dos escores transformados  $z_1$  e  $z_2$

classificados em Saúde, Educação e Renda, que descrevem e interpretam os escores. A coleção dos dados foi realizada pelo IPE (Instituto Peruano de Economia) e pelo INEI (Instituto Nacional de Estadística e Informática). Para mais detalhes, consulte as informações disponíveis em [INEI \(2019\)](#). A variável educação considera três tipos de indicadores para uma pessoa: educação completa na escola, média de anos em educação e sucesso educacional. Por fim, propomos um modelo de regressão Beta para cada variável resposta, utilizando o pacote GAMLSS ([STASINOPOULOS et al., 2017](#)) a fim de explicar os escores como uma função dos indicadores de desenvolvimento humano em cada região peruana.

Tabela 11 – Indicadores de desenvolvimento humano do Peru

|    | Região        | Saúde | Educação 1 | Educação 2 | Educação 3 | Renda |
|----|---------------|-------|------------|------------|------------|-------|
| 1  | Amazonas      | 0,733 | 0,455      | 0,329      | 0,387      | 0,196 |
| 2  | Áncash        | 0,832 | 0,632      | 0,438      | 0,526      | 0,314 |
| 3  | Apurímac      | 0,746 | 0,655      | 0,346      | 0,476      | 0,257 |
| 4  | Arequipa      | 0,877 | 0,759      | 0,580      | 0,664      | 0,456 |
| 5  | Ayacucho      | 0,803 | 0,629      | 0,361      | 0,476      | 0,212 |
| 6  | Cajamarca     | 0,805 | 0,512      | 0,315      | 0,402      | 0,238 |
| 7  | Callao        | 0,885 | 0,731      | 0,575      | 0,648      | 0,457 |
| 8  | Cusco         | 0,790 | 0,718      | 0,459      | 0,574      | 0,296 |
| 9  | Huancavelica  | 0,820 | 0,586      | 0,298      | 0,418      | 0,165 |
| 10 | Huánuco       | 0,792 | 0,578      | 0,369      | 0,462      | 0,255 |
| 11 | Ica           | 0,864 | 0,735      | 0,580      | 0,653      | 0,383 |
| 12 | Junín         | 0,799 | 0,673      | 0,481      | 0,569      | 0,293 |
| 13 | La Libertad   | 0,865 | 0,608      | 0,459      | 0,528      | 0,361 |
| 14 | Lambayeque    | 0,876 | 0,693      | 0,472      | 0,572      | 0,305 |
| 15 | Lima          | 0,878 | 0,755      | 0,612      | 0,680      | 0,593 |
| 16 | Loreto        | 0,825 | 0,440      | 0,475      | 0,457      | 0,302 |
| 17 | Madre De Dios | 0,825 | 0,640      | 0,494      | 0,562      | 0,498 |
| 18 | Moquegua      | 0,852 | 0,739      | 0,566      | 0,647      | 0,520 |
| 19 | Pasco         | 0,802 | 0,678      | 0,457      | 0,557      | 0,245 |
| 20 | Piura         | 0,868 | 0,624      | 0,431      | 0,519      | 0,300 |
| 21 | Puno          | 0,819 | 0,738      | 0,420      | 0,557      | 0,221 |
| 22 | San Martín    | 0,767 | 0,527      | 0,384      | 0,450      | 0,327 |
| 23 | Tacna         | 0,831 | 0,728      | 0,559      | 0,638      | 0,388 |
| 24 | Tumbes        | 0,794 | 0,681      | 0,502      | 0,585      | 0,369 |
| 25 | Ucayali       | 0,761 | 0,495      | 0,464      | 0,479      | 0,310 |

*Explicando o escore transformado da componente  $x_1$  (preferência por partidos progressistas)*

O escore transformado para a componente  $x_1$  é representado como  $z_{i1}$ , e varia entre zero e um. Sejam  $z_{11}, \dots, z_{251}$  variáveis aleatórias independentes, onde cada  $z_{i1}$ ,  $i = 1, \dots, 25$ , segue uma distribuição Beta com média  $\mu_{i1}$  e parâmetro de precisão desconhecido  $\phi_1$ . O modelo é

obtido considerando a função de ligação logito, e a média  $\mu_{i1}$  é representada da seguinte forma:

$$g(\mu_{i1}) = \log\left(\frac{\mu_{i1}}{1 - \mu_{i1}}\right) = \mathbf{U}_i^\top \boldsymbol{\beta}_1 = \beta_1 + \beta_2 \text{Saúde}_i + \beta_3 \text{Educação1}_i + \beta_4 \text{Educação2}_i + \beta_5 \text{Educação3}_i + \beta_6 \text{Renda}_i$$

em que  $\mathbf{U}_i = (1, u_{i1}, \dots, u_{i5})$  representa as observações das covariáveis, e  $\boldsymbol{\beta}_1 = (\beta_1, \dots, \beta_6)^\top$  é um vetor de parâmetros de regressão desconhecidos ( $\boldsymbol{\beta}_1 \in \mathbb{R}^6$ ).

As covariáveis incluem indicadores de desenvolvimento humano, saúde, educação e renda. No entanto, ao estimarmos a correlação entre as covariáveis Educação 2 e Educação 3, encontramos um valor de 0,93, indicando uma alta correlação entre essas covariáveis. Consequentemente, excluímos a covariável Educação 3 do modelo porque o número médio de anos de estudo é considerado mais importante do que o sucesso educacional.

Para avaliar a significância das variáveis no modelo, examinamos os p-valores. A [Tabela 12](#) apresenta os resultados do modelo de regressão utilizando o pacote GAMLSS. Notavelmente, a variável Educação 1 é estatisticamente significativa, com um p-valor de 0,011. Além disso, seu coeficiente revela um impacto positivo na variável resposta, com um valor de 9,56.

Tabela 12 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente  $x_1$

|                          | Estimate | Std. Error | t value | Pr(> t ) |
|--------------------------|----------|------------|---------|----------|
| Intercept.               | 3,769    | 4,430      | 0,851   | 0,405    |
| Saúde                    | -6,946   | 6,492      | -1,070  | 0,298    |
| Educação1                | 9,567    | 3,413      | 2,803   | 0,011 *  |
| Educação2                | -4,866   | 5,793      | -0,840  | 0,411    |
| Renda                    | -6,883   | 3,965      | -1,736  | 0,098    |
| logito( $\hat{\phi}_1$ ) | -0,089   | 0,196      | -0,457  | 0,653    |

*Nível de significância:  $p \leq 0,001$  '\*\*\*';  $0,001 < p \leq 0,01$  '\*\*';  $0,01 < p \leq 0,05$  '\*'*

Em seguida, elaboramos outro modelo excluindo a variável Educação 2 devido ao seu p-valor mais elevado, indicando falta de significância. Os resultados em [Tabela 13](#) destacam que Educação 1 e Renda são as únicas variáveis significativas, com p-valores de 0,007 e 0,000, respectivamente. Analisando os coeficientes, observamos que Educação 1 possui um impacto positivo de 8,09, enquanto Renda exerce um impacto negativo de -9,63 na variável resposta.

Finalmente, elaboramos outro modelo de regressão considerando exclusivamente essas duas variáveis significativas (Educação 1 e Renda). Os resultados na [Tabela 14](#) mostram o p-valor de 0,011 e 0,000 respectivamente, e os coeficientes indicam que cada unidade adicional na variável Educação 1 (pessoa com escola completa) está associado com um aumento de 6,77 na preferência por partidos progressistas, e cada unidade adicional em Renda está associado a uma diminuição de 10,79 na preferência por partidos progressistas.

Tabela 13 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente  $x_1$  sem a variável Educação 2

|                          | Estimate | Std. Error | t value | Pr(> t ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | 3,758    | 4,386      | 0,857   | 0,401    |     |
| Saúde                    | -7,373   | 6,385      | -1,155  | 0,261    |     |
| Educação1                | 8,090    | 2,714      | 2,981   | 0,007    | **  |
| Renda                    | -9,638   | 2,423      | -3,979  | 0,000    | *** |
| logito( $\hat{\phi}_1$ ) | -0,061   | 0,195      | -0,316  | 0,755    |     |

*Nível de significância:  $p \leq 0,001$  '\*\*\*';  $0,001 < p \leq 0,01$  '\*\*';  $0,01 < p \leq 0,05$  '\*'*

Tabela 14 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente  $x_1$  sem as variáveis Educação 2 e Saúde

|                          | Estimate | Std. Error | t value | Pr(> t ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | -1,073   | 1,401      | -0,766  | 0,452    |     |
| Educação1                | 6,770    | 2,455      | 2,757   | 0,011    | **  |
| Renda                    | -10,785  | 2,272      | -4,747  | 0,000    | *** |
| logito( $\hat{\phi}_1$ ) | -0,022   | 0,196      | -0,117  | 0,908    |     |

*Nível de significância:  $p \leq 0,001$  '\*\*\*';  $0,001 < p \leq 0,01$  '\*\*';  $0,01 < p \leq 0,05$  '\*'*

### Explicando o escore transformado da componente $x_2$ ( preferência por partidos tradicionais)

O escore transformado para a componente  $x_2$  é representado como  $z_{i2}$ , e varia entre zero e um.. Sejam  $z_{12}, \dots, z_{252}$  variáveis aleatórias independentes, onde cada  $z_{i2}$ ,  $i = 1, \dots, 25$ , segue uma distribuição Beta com média  $\mu_{i2}$  e parâmetro de precisão desconhecido  $\phi_2$ . O modelo é obtido considerando a função de ligação logito, e a média  $\mu_{i2}$  é representada como:

$$g(\mu_{i2}) = \log\left(\frac{\mu_{i2}}{1 - \mu_{i2}}\right) = \mathbf{U}_i^\top \boldsymbol{\beta}_2 = \beta_1 + \beta_2 \text{Saúde}_i + \beta_3 \text{Educação1}_i + \beta_4 \text{Educação2}_i + \beta_5 \text{Educação3}_i + \beta_6 \text{Renda}_i$$

em que  $\mathbf{U}_i = (1, u_{i1}, \dots, u_{i5})$  representa as observações das covariáveis, e  $\boldsymbol{\beta}_2 = (\beta_1, \dots, \beta_6)^\top$  é um vetor de parâmetros de regressão desconhecidos ( $\boldsymbol{\beta}_2 \in \mathbb{R}^6$ ).

A Tabela 15 apresenta os resultados do modelo de regressão utilizando o pacote GAMLSS, destacando que a variável Education 1 é a única significativa. Notavelmente, o p-valor associado a Educação 1 é próximo de zero com um valor de 0.002. Além disso, o coeficiente do parâmetro indica um impacto negativo de -9,36 na variável resposta.

Seguidamente, elaboramos outro modelo excluindo a variável Saúde devido ao seu p-valor mais elevado, e indicando que não é significativa no modelo. A Tabela 16 mostra os resultados, onde Educação 1 é novamente a única significativa com um p-valor de 0,001.

Por último, observamos que os dois primeiros modelos têm como única variável significativa Educação 1, então desenvolvemos outro modelo de regressão considerando apenas essa variável significativa. Como os escores transformados da componente  $x_2$  indicam uma preferência

Tabela 15 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente  $x_2$ 

|                          | Estimate | Std. Error | t value | Pr(> t ) |    |
|--------------------------|----------|------------|---------|----------|----|
| Intercept.               | 8,559    | 4,117      | 2,079   | 0,051    | .  |
| Saúde                    | -0,405   | 6,003      | -0,067  | 0,946    |    |
| Educação1                | -9,363   | 2,704      | -3,463  | 0,002    | ** |
| Educação2                | -2,964   | 4,901      | -0,605  | 0,552    |    |
| Renda                    | -2,743   | 3,777      | -0,726  | 0,476    |    |
| logito( $\hat{\phi}_2$ ) | -0,229   | 0,195      | -1,174  | 0,255    |    |

*Nível de significância:  $p \leq 0,001$  '\*\*\*';  $0,001 < p \leq 0,01$  '\*\*';  $0,01 < p \leq 0,05$  '\*'*

Tabela 16 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente  $x_2$  sem a variável Saúde

|                          | Estimate | Std. Error | t value | Pr(> t ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | 8.299    | 1.464      | 5.669   | 0,000    | *** |
| Educação1                | -9,411   | 2,608      | -3,609  | 0,001    | **  |
| Educação2                | -3,024   | 4,825      | -0,627  | 0,537    |     |
| Renda                    | -2,781   | 3,735      | -0,745  | 0,465    |     |
| logito( $\hat{\phi}_2$ ) | -0,229   | 0,195      | -1,174  | 0,254    |     |

*Nível de significância:  $p \leq 0,001$  '\*\*\*';  $0,001 < p \leq 0,01$  '\*\*';  $0,01 < p \leq 0,05$  '\*'*

por partidos tradicionais, os resultados na [Tabela 17](#) mostram que cada unidade adicional na variável Educação 1 está associado a uma diminuição de 12,7 ao escolher um partido político tradicional.

Tabela 17 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente  $x_2$  considerando unicamente covariáveis significativas

|                          | Estimate | Std. Error | t value | Pr(> t ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | 8,096    | 1,500      | 5,398   | 0,000    | *** |
| Educação1                | -12,703  | 2,294      | -5,538  | 0,000    | *** |
| logito( $\hat{\phi}_2$ ) | -0,095   | 0,197      | -0,483  | 0,634    |     |

*Nível de significância:  $p \leq 0,001$  '\*\*\*';  $0,001 < p \leq 0,01$  '\*\*';  $0,01 < p \leq 0,05$  '\*'*

### 3.3.6 Diagnóstico dos resíduos

Conforme os resíduos quantílicos normalizados proposto por [Dunn e Smyth \(1996\)](#), Para obter os resultados da análise residual, utilizamos o pacote GAMLSS mediante o comando "plot". Em seguida, é mostrado quatro gráficos: Valores ajustados vs. Resíduos, Índice vs. Resíduos, Densidade e QQplot dos resíduos do escore transformado 1 e 2, para verificar a homocedasticidade e identificar pontos influentes.

Na [Figura 8](#), os gráficos de Valores ajustados vs. Resíduos (a) e Índice vs. Resíduos (b) sugerem homocedasticidade, uma vez que os resíduos estão dispersos ao redor de zero e não exibem padrões discerníveis, indicando a presença de linearidade. Além disso, o QQplot

mostra que os resíduos variam entre -3 e 3, não revelando pontos influentes, pois estão dentro do intervalo de confiança.

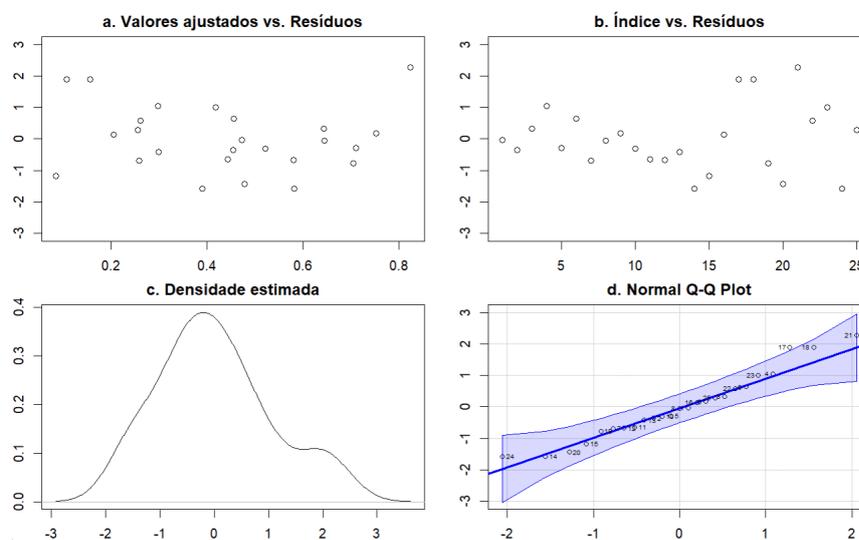


Figura 8 – Diagnóstico de resíduos para  $z_1$  no Peru

Na Figura 9, Valores ajustados vs. Resíduos (a) e Índice vs. Resíduos (b) indicam que não há um padrão observável e atendem à homocedasticidade, onde os pontos residuais estão dispersos homogeneamente. No entanto, o Q-Q Normal em (d) mostra a região de Tacna como um ponto influente nos resíduos, pois está fora do intervalo de confiança.

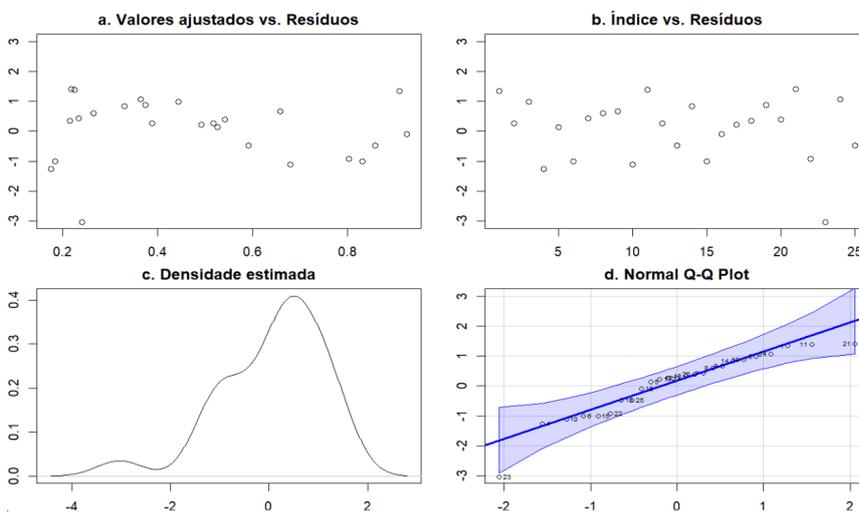


Figura 9 – Diagnóstico de resíduos para  $z_2$  no Peru

### 3.4 Discussão final

Em síntese, aplicamos a metodologia proposta aos dados coletados das eleições no Peru, conforme descrito na etapa 1. Esses dados foram transformados em proporções na etapa 2 e, devido às características composicionais e suas restrições inerentes, foram posteriormente

convertidos em razões logarítmicas na etapa 3. Na etapa 4, identificamos duas componentes principais que retêm a maioria da variância dos dados. A componente principal  $x_1$  indica a preferência por partidos progressistas, enquanto a componente principal  $x_2$  indica a preferência por partidos tradicionais. O espaço de componentes também revelou as cargas das variáveis em três direções distintas onde as cargas de ACCP e PL mostraram uma correlação positiva com  $x_1$ , indicando associação com a preferência por partidos progressistas, enquanto FP mostrou uma correlação negativa, sugerindo o oposto. Da mesma forma, FP mostrou uma correlação positiva com  $x_2$ , enquanto RP e AVP mostraram uma correlação negativa, indicando associação com a preferência por partidos tradicionais.

Na etapa 5, obtivemos os escores das regiões, apresentando variabilidade distinta em cada componente. Portanto, na etapa 6, normalizamos os dados para o intervalo unitário.

Dando continuidade à análise, na etapa 7, coletamos dados que descrevem o comportamento da sociedade peruana associado à eleição de um candidato de governo. Na etapa 8, empregamos o pacote GAMLSS para obter as estimativas do modelo de regressão Beta nas eleições presidenciais peruanas de 2021. Os resultados indicam que, para a componente  $x_1$ , a variável Educação 1 possui um coeficiente positivo, sugerindo que regiões com maior taxa de conclusão escolar tendem a preferir partidos progressistas. Por outro lado, a variável Renda exibe um coeficiente negativo, indicando que regiões com renda mais alta têm menor propensão a preferir partidos progressistas. Na componente  $x_2$ , a variável Educação 1 foi a única significativa, apresentando um coeficiente negativo. Isso sugere que regiões com maior nível de conclusão escolar tendem a não escolher partidos tradicionais, possivelmente devido ao desempenho insatisfatório desses partidos no governo ou a escândalos de corrupção envolvendo seus líderes.

Finalmente, conduzimos um diagnóstico dos resíduos normalizados para avaliar se o modelo de regressão Beta proposto se ajusta adequadamente aos dados. Os gráficos para o escore 1 e escore 2 transformados revelaram resíduos sem padrões aparentes, distribuídos homoganeamente. Além disso, o QQplot indicou que os resíduos se aproximam de uma distribuição normal, mantendo-se dentro dos limites com 95% de confiança.

---

# UMA ABORDAGEM ESTATÍSTICA PARA A ANÁLISE DAS ELEIÇÕES PRESIDENCIAIS DO BRASIL 2022

---

---

## 4.1 Introdução

As eleições brasileiras ocorrem a cada quatro anos em 27 Unidades da Federação, sendo 26 estados e o Distrito Federal. São escolhidos o presidente, governadores, senadores, deputados federais e estaduais. Quando nenhum dos candidatos atinge 50% mais um por cento dos votos, realiza-se um segundo turno entre os dois primeiros candidatos que obtiveram a maioria dos votos.

O primeiro turno das eleições Brasileiras de 2022 ocorreram em 2 de outubro, onde os partidos que obtiveram a maioria de votos foram: Partido dos trabalhadores (PT) com 48,43%, Partido Liberal(PL) com 43,2%, Partido democratico trabalhista(PDT) com 3,04%, Movimento Democratico Brasileiro(MDB) com 4.16%, União com 0,51% e Novo com 0,47%. Assim, observamos que os candidatos do Partido dos Trabalhadores e do Partido Liberal conquistaram a maioria dos votos. Em seguida, os dois candidatos participaram do segundo turno em 30 de outubro de 2022, onde o Partido Trabalhista obteve 50,9% dos votos e o Partido Liberal, 49,1%. Por fim, o Tribunal Superior Eleitoral foi o órgão responsável por reconhecer o candidato com a maior quantidade de votos da população.

Para uma análise apropriada, consideramos os partidos participantes designados da seguinte maneira: PT, PL, MDB, PDT e os outros partidos, como UNIÃO e NOVO, foram agrupados sob a variável "OUTROS" devido à sua baixa porcentagem de votos. Os dados referentes aos votos no primeiro turno das eleições brasileiras estão apresentados na [Tabela 18](#), obtida do [TSE \(2022\)](#).

Tabela 18 – Dados das eleições de Brasil 2022

|    | UF | Territorialidade    | PT         | PL         | MDB       | PDT     | OUTROS    | Total      |
|----|----|---------------------|------------|------------|-----------|---------|-----------|------------|
| 1  | AC | Acre                | 129.022    | 275.582    | 20.122    | 12.314  | 18.863    | 455.903    |
| 2  | AL | Alagoas             | 974.156    | 621.515    | 67.411    | 43.542  | 99.347    | 1.805.971  |
| 3  | AP | Amapá               | 197.382    | 187.621    | 27.497    | 14.670  | 15.672    | 442.842    |
| 4  | AM | Amazonas            | 1.019.684  | 880.198    | 87.060    | 44.527  | 82.302    | 2.113.771  |
| 5  | BA | Bahia               | 5.873.081  | 2.047.599  | 197.305   | 217.224 | 539.632   | 8.874.841  |
| 6  | CE | Ceará               | 3.578.355  | 1.377.827  | 66.214    | 369.222 | 236.992   | 5.628.610  |
| 7  | DF | Distrito Federal    | 649.534    | 910.397    | 105.377   | 74.308  | 80.284    | 1.819.900  |
| 8  | ES | Espírito Santo      | 897.348    | 1.160.030  | 85.325    | 56.221  | 116.965   | 2.315.889  |
| 9  | GO | Goiás               | 1.454.723  | 1.920.203  | 170.742   | 90.695  | 176.234   | 3.812.597  |
| 10 | MA | Maranhão            | 2.603.454  | 983.861    | 78.254    | 96.095  | 158.771   | 3.920.435  |
| 11 | MT | Mato Grosso         | 633.748    | 1.102.866  | 55.989    | 29.437  | 70.140    | 1.892.180  |
| 12 | MS | Mato Grosso do Sul  | 588.323    | 794.206    | 79.719    | 29.314  | 63.587    | 1.555.149  |
| 13 | MG | Minas Gerais        | 5.802.571  | 5.239.264  | 500.658   | 310.324 | 802.411   | 12.655.228 |
| 14 | PA | Pará                | 2.443.730  | 1.884.673  | 204.075   | 116.057 | 140.776   | 4.789.311  |
| 15 | PB | Paraíba             | 1.554.868  | 717.416    | 57.154    | 76.225  | 151.816   | 2.557.479  |
| 16 | PR | Paraná              | 2.363.492  | 3.628.612  | 309.685   | 180.599 | 346.155   | 6.828.543  |
| 17 | PE | Pernambuco          | 3.558.322  | 1.630.938  | 96.570    | 130.015 | 322.526   | 5.738.371  |
| 18 | PI | Piauí               | 1.518.008  | 406.897    | 42.179    | 59.321  | 89.240    | 2.115.645  |
| 19 | RJ | Rio de Janeiro      | 3.847.143  | 4.831.246  | 365.969   | 301.489 | 563.616   | 9.909.463  |
| 20 | RN | Rio Grande do Norte | 1.264.179  | 622.731    | 38.633    | 71.740  | 93.321    | 2.090.604  |
| 21 | RS | Rio Grande do Sul   | 2.806.672  | 3.245.023  | 317.957   | 190.945 | 329.419   | 6.890.016  |
| 22 | RO | Rondônia            | 261.749    | 581.306    | 31.217    | 19.353  | 33.202    | 926.827    |
| 23 | RR | Roraima             | 68.760     | 207.587    | 12.956    | 6.709   | 9.392     | 305.404    |
| 24 | SC | Santa Catarina      | 1.279.216  | 2.694.406  | 191.310   | 88.672  | 233.870   | 4.487.474  |
| 25 | SP | São Paulo           | 10.490.032 | 12.239.989 | 1.625.596 | 898.540 | 1.935.557 | 27.189.714 |
| 26 | SE | Sergipe             | 828.716    | 378.610    | 42.073    | 40.247  | 75.078    | 1.364.724  |
| 27 | TO | Tocantins           | 434.303    | 379.194    | 25.209    | 18.141  | 34.602    | 891.449    |

PT: Partido dos Trabalhadores, PL: Partido Liberal, MDB: Movimento Democrático Brasileiro, PDT: Partido Democrático Trabalhista, e OUTROS

Em seguida, podemos observar que as regiões com o maior número de votos são: São Paulo (27.189.714), Minas Gerais (12.655.228) e Rio de Janeiro (9.909.463). Já as regiões com o menor número de votos são: Roraima (305.404), Amapá (442.842) e Acre (455.903).

## 4.2 Metodologia

Dado que buscamos interpretar os votos dos estados do Brasil, aplicamos a metodologia proposta nas eleições presidenciais do ano de 2022 considerando as seguintes etapas.

1. Coleta de dados das eleições presidenciais de Brasil do ano de 2022.
2. Transformação dos votos do Brasil em proporções.
3. Aplicação da transformação em razões logarítmicas das proporções de votos do Brasil.
4. Aplicação da análise de componentes principais (ACP) para dados composicionais transformados do Brasil.
5. Identificação de escores das regiões de Brasil por meio da ACP.

6. Transformação dos escores para o intervalo unitário.
7. Coleta de indicadores de desenvolvimento humano (IDH) do Brasil.
8. Aplicação de um modelo de regressão Beta à variável escore transformada.

Na etapa 1, realizamos a coleta de dados dos votos no Brasil, evidenciando variabilidades distintas por estado. Na etapa 2, realizamos a transformação dos dados dos votos em proporções  $V_{ij}$ . Dada a natureza composicional dos dados, na etapa 3, aplicamos a transformação de razões logarítmicas  $V_{ij}^*$  para eliminar as restrições.

Em seguida, na etapa 4, realizamos a análise de componentes principais para obter as componentes que contêm a maior informação dos dados. Na etapa 5, obtivemos os escores  $x_{ij}$  dos estados do Brasil. Na etapa 6, os escores foram transformados em  $z_{ij}$  com valores entre zero e um. Considerando a interpretação das componentes principais por meio dos escores, na etapa 7, incorporamos os indicadores de desenvolvimento humano do Brasil, classificados em longevidade, educação e renda. Finalmente, na etapa 8, aplicamos o modelo de regressão Beta, considerando os escores como variável resposta e os indicadores de desenvolvimento humano do Brasil como covariáveis. A Figura 10 apresenta o diagrama com detalhes sobre o processo.

Os códigos da metodologia proposta foram desenvolvidos no software R e estão disponíveis na [Código-fonte 2](#). Além disso, podem ser acessado pelo seguinte link: [Bounded regression model Brasil 2022](#).

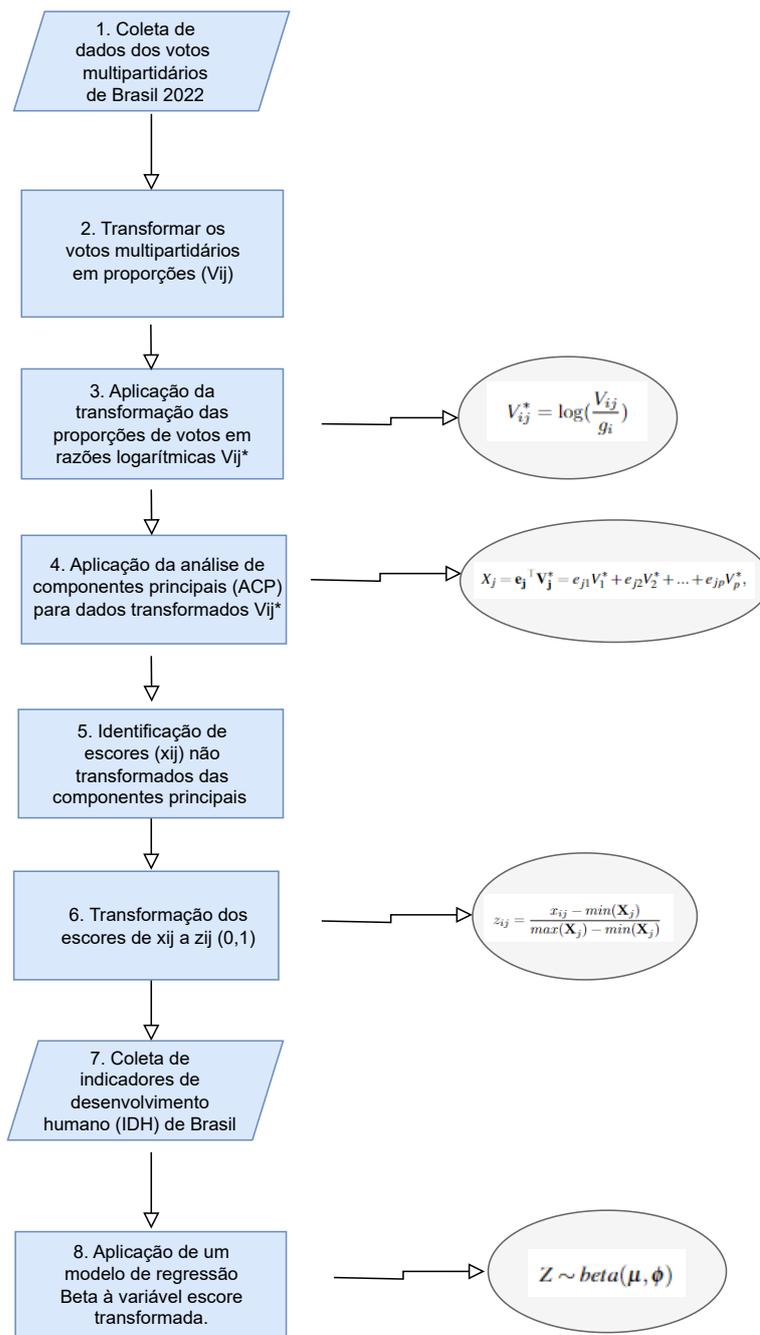


Figura 10 – Diagrama da metodologia nas eleições Brasileiras 2022

## 4.3 Resultados

### 4.3.1 Matriz de correlação das eleições Brasileiras de 2022

A Tabela 19 apresenta a correlação dos dados transformados clr. Como foi discutido anteriormente, os valores são apropriados para análise e não há correlação espúria porque as variáveis são independentes e a restrição de soma constante foi eliminada pela transformação. Além disso, os valores indicam correlações negativas e positivas entre as variáveis. Para exemplificar, os partidos políticos PT e PL exibem uma correlação negativa significativa de -0,83, assim como

os partidos PT e MDB, que apresentam uma correlação de 0,79.

Tabela 19 – Matriz de correlação considerando as transformações clr dos votos nas eleições no Brasil em 2022

|        | PT     | PL     | MDB    | PDT    | OUTROS |
|--------|--------|--------|--------|--------|--------|
| PT     | 1,000  | -0,828 | -0,792 | 0,427  | 0,402  |
| PL     | -0,828 | 1,000  | 0,636  | -0,589 | -0,501 |
| MDB    | -0,792 | 0,636  | 1,000  | -0,651 | -0,539 |
| PDT    | 0,427  | -0,589 | -0,651 | 1,000  | 0,039  |
| OUTROS | 0,402  | -0,501 | -0,539 | 0,039  | 1,000  |

### 4.3.2 Identificando as principais componentes das eleições Brasileiras

De maneira semelhante ao trabalho realizado com os dados das eleições peruanas de 2021, identificamos as principais componentes da análise de componentes principais (ACP) associadas às eleições brasileiras de 2022. Para conduzir essa análise, utilizamos o pacote R para análise composicional proposto por [Boogaart e Tolosana-Delgado \(2008\)](#). Os resultados revelaram duas métricas essenciais: cargas e escores, além da variância acumulada em cada componente. Em seguida, aplicamos o biplot ([GABRIEL, 1971](#)) para visualizar as cargas dos partidos políticos brasileiros e os escores dos estados de Brasil nas principais componentes.

#### *Análise de componentes principais da transformação clr dos dados*

A [Tabela 20](#) exibe as cargas e a variância acumulada nas componentes para os dados de Brasil com transformação clr. As duas primeiras componentes explicam 84%, conforme estimado pelo pacote `compositions` do R. As cargas observadas são negativas, positivas e próximas de zero, indicadas por espaços vazios. Denotamos a primeira componente com  $x_1$  e a segunda componente com  $x_2$ .

Tabela 20 – Cargas da análise de componentes principais da transformação clr dos dados de Brasil

| Parties         | Comp.1 ( $x_1$ ) | Comp.2 ( $x_2$ ) | Comp.3 | Comp.4 | Comp.5 |
|-----------------|------------------|------------------|--------|--------|--------|
| PT              | 0,494            |                  | 0,601  | -0,346 | -0,523 |
| PL              | -0,499           |                  | -0,364 | -0,626 | -0,475 |
| MDB             | -0,505           |                  | 0,316  | 0,613  | -0,518 |
| PDT             | 0,378            | -0,657           | -0,462 | 0,254  | -0,383 |
| OUTROS          | 0,329            | 0,752            | -0,439 | 0,219  | -0,292 |
| Desvio padrão   | 1,800            | 0,980            | 0,680  | 0,578  | 0,000  |
| Var. proporção  | 0,648            | 0,192            | 0,092  | 0,066  | 0,000  |
| Var. cumulativa | 0,648            | 0,840            | 0,933  | 1,000  | 1,000  |

O biplot mostra os escores dos estados Brasileiros nas duas primeiras componentes, onde visualizamos o PT e o PL, que obtiveram a maior quantidade de votos (representando



### 4.3.4 Transformando os escores da componente principal $x_1$

A Figura 11 exibe a posição ou escore  $x_{i1}$ , em que  $i$  indica os estados do Brasil para a componente  $x_1$ . Como os valores tem diferente variabilidade nas componentes e queremos facilitar a interpretação, realizamos uma transformação monotônica nos escores. Finalmente, obtemos valores normalizados entre 0 e 1 por meio da seguinte transformação:

$$z_{i1} = \frac{x_{i1} - \min(\mathbf{x}_1)}{\max(\mathbf{x}_1) - \min(\mathbf{x}_1)}$$

Aqui,  $0 \leq z_{i1} \leq 1$  representa os escores transformados exclusivamente da componente principal  $x_1$ . O vetor  $\mathbf{x}_1 = (x_{1j}, \dots, x_{nj})^T$  representa os escores não transformados de  $x_1$ , e os estados são denotados por  $i = 1, \dots, 27$ . A Tabela 21 mostra os valores dos escores transformados para a componente  $x_1$  no Brasil.

Tabela 21 – Escores transformados da componente  $x_1$

| UF | $x_1$  | $z_1$ |
|----|--------|-------|
| AC | -1,951 | 0,160 |
| AL | 0,638  | 0,481 |
| AP | -0,913 | 0,289 |
| AM | -0,711 | 0,314 |
| BA | 2,932  | 0,765 |
| CE | 4,833  | 1,000 |
| DF | -1,256 | 0,246 |
| ES | -0,864 | 0,295 |
| GO | -1,267 | 0,245 |
| MA | 2,729  | 0,740 |
| MT | -1,460 | 0,221 |
| MS | -1,823 | 0,176 |
| MG | -0,023 | 0,399 |
| PA | -0,551 | 0,334 |
| PB | 2,541  | 0,716 |
| PR | -1,321 | 0,238 |
| PE | 2,474  | 0,708 |
| PI | 3,330  | 0,814 |
| RJ | -0,335 | 0,360 |
| RN | 2,743  | 0,741 |
| RS | -0,988 | 0,280 |
| RO | -2,077 | 0,145 |
| RR | -3,247 | 0,000 |
| SC | -2,133 | 0,138 |
| SP | -1,194 | 0,254 |
| SE | 1,572  | 0,596 |
| TO | -0,160 | 0,382 |

Seguidamente, a Figura 12 mostra o histograma dos escores transformados da primeira componente principal, com valores variando entre 0 e 1.

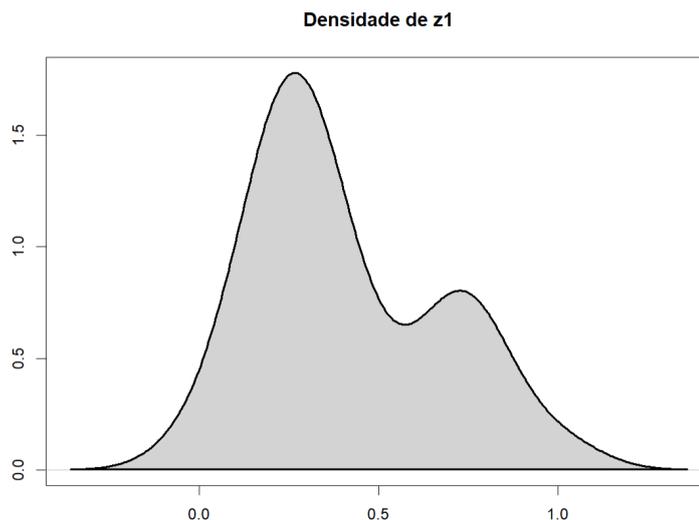


Figura 12 – Histograma do escore transformado  $z_1$

### 4.3.5 Proposta do modelo de regressão Beta

Posteriormente, considerando os escores transformados entre 0 e 1, sugerimos um modelo de regressão de resposta limitada. Os dados foram coletados pela AtlasBR (2022). Eles estão detalhados na Tabela 22, que inclui os principais indicadores de desenvolvimento humano relacionados ao Brasil, sendo classificados em Renda, Educação e Longevidade, juntamente com as respectivas unidades federativas.

*Explicando o escore 1 da componente  $x_1$  (preferência por partidos progressistas)*

O escore transformado para a componente  $x_1$  dos dados de Brasil é representada como  $z_{i1}$ , variando entre zero e um. Sejam  $z_{11}, \dots, z_{271}$  variáveis aleatórias independentes, onde cada  $z_{i1}$ ,  $i = 1, \dots, 27$  segue uma distribuição Beta com media  $\mu_{i1}$  e parâmetro de precisão desconhecido  $\phi_1$ . O modelo é obtido considerando a função de ligação logito, e a média  $\mu_{i1}$  é representada da seguinte forma:

$$g(\mu_{i1}) = \log \frac{\mu_{i1}}{1 - \mu_{i1}} = \mathbf{U}_i^\top \boldsymbol{\beta}_1 = \beta_1 + \beta_2 \text{Saúde}_i + \beta_3 \text{Educação}_i + \beta_4 \text{Longevidade}_i$$

em que  $\mathbf{U}_i = (1, u_{i1}, u_{i2}, u_{i3})$  representa as observações das covariáveis, e  $\boldsymbol{\beta}_1 = (\beta_1, \dots, \beta_4)^\top$  é um vetor de parâmetros de regressão desconhecidos ( $\boldsymbol{\beta}_1 \in \mathbb{R}^4$ ).

A Tabela 23 exibe os resultados da regressão utilizando o pacote GAMLSS. Nota-se que as variáveis Renda e Educação são estatisticamente significativas, pois seus p-valores estão próximos de zero com 0,000 e 0,002 respectivamente, enquanto a variável Longevidade apresenta um p-valor superior a 0,05. Além disso, os coeficientes dos parâmetros revelam um impacto positivo de 19,68 de Educação e um impacto negativo de -27,39 de Renda na variável resposta.

Em seguida, na Tabela 24, desenvolvemos outro modelo de regressão considerando

Tabela 22 – Indicadores de desenvolvimento humano do Brasil

|    | UF | Territorialidade    | Renda | Educação | Longevidade |
|----|----|---------------------|-------|----------|-------------|
| 1  | AC | Acre                | 0,655 | 0,692    | 0,788       |
| 2  | AL | Alagoas             | 0,630 | 0,679    | 0,748       |
| 3  | AP | Amapá               | 0,648 | 0,647    | 0,778       |
| 4  | AM | Amazonas            | 0,641 | 0,720    | 0,744       |
| 5  | BA | Bahia               | 0,648 | 0,659    | 0,772       |
| 6  | CE | Ceará               | 0,658 | 0,766    | 0,784       |
| 7  | DF | Distrito Federal    | 0,821 | 0,817    | 0,803       |
| 8  | ES | Espírito Santo      | 0,715 | 0,742    | 0,864       |
| 9  | GO | Goiás               | 0,715 | 0,742    | 0,721       |
| 10 | MA | Maranhão            | 0,603 | 0,716    | 0,715       |
| 11 | MT | Mato Grosso         | 0,720 | 0,758    | 0,730       |
| 12 | MS | Mato Grosso do Sul  | 0,733 | 0,741    | 0,751       |
| 13 | MG | Minas Gerais        | 0,718 | 0,762    | 0,846       |
| 14 | PA | Pará                | 0,645 | 0,686    | 0,744       |
| 15 | PB | Paraíba             | 0,653 | 0,669    | 0,779       |
| 16 | PR | Paraná              | 0,744 | 0,780    | 0,785       |
| 17 | PE | Pernambuco          | 0,647 | 0,721    | 0,797       |
| 18 | PI | Piauí               | 0,649 | 0,698    | 0,726       |
| 19 | RJ | Rio de Janeiro      | 0,759 | 0,758    | 0,769       |
| 20 | RN | Rio Grande do Norte | 0,69  | 0,68     | 0,82        |
| 21 | RS | Rio Grande do Sul   | 0,767 | 0,750    | 0,797       |
| 22 | RO | Rondônia            | 0,677 | 0,694    | 0,731       |
| 23 | RR | Roraima             | 0,680 | 0,673    | 0,745       |
| 24 | SC | Santa Catarina      | 0,759 | 0,790    | 0,827       |
| 25 | SP | São Paulo           | 0,771 | 0,839    | 0,810       |
| 26 | SE | Sergipe             | 0,662 | 0,684    | 0,764       |
| 27 | TO | Tocantins           | 0,684 | 0,732    | 0,779       |

Tabela 23 – Estimativas dos parâmetros do modelo de regressão Beta para o escore transformado da componente  $x_1$  no Brasil

|                          | Estimate | Std. Error | t value | Pr(> t ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | -3,761   | 4,366      | -0,861  | 0,398    |     |
| Renda                    | -27,391  | 6,627      | -4,133  | 0,000    | *** |
| Educação                 | 19,676   | 5,770      | 3,410   | 0,002    | **  |
| Longevidade              | 10,437   | 5,974      | 1,747   | 0,094    | .   |
| logito( $\hat{\phi}_1$ ) | 0,023    | 0,189      | 0,125   | 0,902    |     |

Nível de significância:  $p \leq 0,001$  \*\*\*;  $0,001 < p \leq 0,01$  \*\*;  $0,01 < p \leq 0,05$  \*

exclusivamente com essas duas variáveis significativas. Observa-se que, dado que a componente principal  $x_1$  indica a preferência por partidos progressistas, e cada unidade adicional na variável Renda está associado a uma diminuição de 24,35 na preferência por partidos progressistas, enquanto cada unidade adicional em Educação está associado a um aumento de 20,85 na preferência pelos partidos progressistas.

Tabela 24 – Estimativas dos parâmetros do modelo de regressão Beta sem a variável Longevidade

|                          | Estimate | Std. Error | t value | Pr(> t ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | 1,370    | 3,239      | 0,423   | 0,676    |     |
| Renda                    | -24,352  | 6,630      | -3,673  | 0,001    | *** |
| Educação                 | 20,853   | 6,149      | 3,391   | 0,002    | **  |
| logito( $\hat{\phi}_1$ ) | 0,101    | 0,189      | 0,534   | 0,598    |     |

Nível de significância:  $p \leq 0,001$  \*\*\*;  $0,001 < p \leq 0,01$  \*\*;  $0,01 < p \leq 0,05$  \*

### 4.3.6 Diagnóstico dos resíduos

Na Figura 13, os gráficos de Valores ajustados vs. Resíduos (a) e Índice vs. Resíduos (b) indicam homocedasticidade, uma vez que os resíduos não apresentam nenhum padrão discernível em sua dispersão, além disso valores estão distribuídos entre -3 e 2. O gráfico Q-Q Normal (d) revela que todos os resíduos quantílicos estão dentro de limites com 95% de confiança, embora a densidade mostre uma leve deformação no lado esquerdo devido à baixa probabilidade de ocorrência de resíduos quantílicos entre -4 e -2.

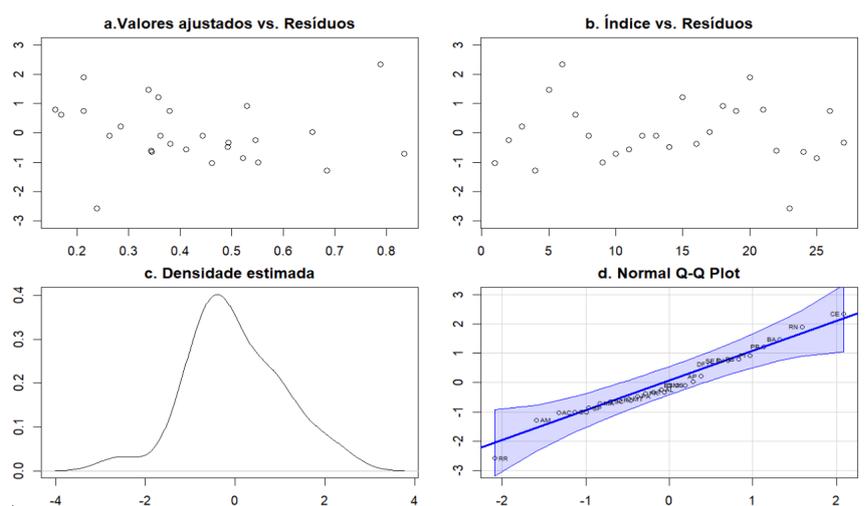


Figura 13 – Diagnóstico de resíduos para  $z_1$  no Brasil

## 4.4 Discussão final

Em síntese, aplicamos a metodologia proposta aos dados coletados das eleições no Brasil, conforme descrito na etapa 1. Esses dados foram transformados em proporções na etapa 2 e, devido às características composicionais e suas restrições inerentes, foram posteriormente convertidos em razões logarítmicas na etapa 3. Na etapa 4, identificamos duas componentes principais que retêm a maioria da variância dos dados. No entanto, observou-se que a maioria dos votos concentrou-se nos partidos PT e PL. Assim, com base no biplot, a decisão foi focar a análise apenas na componente principal  $x_1$ , considerada crucial para indicar a preferência por partidos progressistas. Posteriormente, ao observar o espaço dimensional das componentes, notamos que as variáveis PT e PDT apresentaram correlação positiva com  $x_1$ , enquanto PL

e MDB mostraram correlação negativa. Essas associações indicam padrões interessantes nas preferências políticas dos estados brasileiros, considerando a ideologia política.

Na etapa 5, obtivemos os escores das regiões, apresentando variabilidade distinta em cada componente. Portanto, na etapa 6, normalizamos os dados para o intervalo unitário.

Para dar continuidade à análise, na etapa 7, coletamos dados de indicadores que descrevem o comportamento da sociedade Brasileira associado à eleição de um candidato de governo. Na etapa 8, empregamos o pacote GAMLSS para obter os valores das estimativas no modelo de regressão Beta, aplicado aos dados das eleições brasileiras de 2022. Os resultados revelam que apenas a variável Educação 1 demonstra significância estatística, exibindo uma estimativa positiva. Isso sugere que estados brasileiros com níveis mais elevados de educação têm uma inclinação para preferir partidos progressistas.

Seguidamente, a ao realizar o diagnóstico dos resíduos, observamos se o modelo de regressão Beta proposto ajustou-se adequadamente aos dados brasileiros. Os gráficos mostraram que resíduos não tem um padrão e eles estão dispersos de forma homogênea. Além disso, o QQplot, indicou que os resíduos normais quantílicos estão dentro dos limites com 95% de confiança.

Finalmente, a análise também sugere que o comportamento político da população do Brasil e do Peru em relação à preferência por partidos progressistas é semelhante, uma vez que a componente  $x_1$  analisa a preferência por partidos progressistas. Um grupo de circunscrições demonstra uma tendência à preferência por partidos progressistas, enquanto outro grupo mostra preferência por partidos conservadores. Além disso, é viável ajustar as etapas e incorporar indicadores adicionais como covariáveis para os diferentes estados do Brasil, levando em consideração os diversos fatores que influenciam a decisão dos votos, e podemos explorar outros modelos de regressão com resposta limitada, a fim de identificar qual se adequa melhor aos dados disponíveis.

---

# MODELO DE REGRESSÃO BIVARIADO VIA CÓPULAS: ELEIÇÕES DE PERU

---

## 5.1 Introdução

Nesta seção, formulamos um modelo de regressão bivariada utilizando cópulas para analisar as eleições presidenciais do Peru em 2021. Conforme descrito nos modelos univariados, após realizar a análise de componentes principais, obtivemos duas variáveis: a preferência pelos partidos progressistas  $Z_1$  e a preferência pelos partidos tradicionais  $Z_2$ , assim como as variáveis explicativas associadas incluem Educação, Saúde e Renda.

Além de examinar os efeitos das variáveis explicativas e a independência das variáveis resposta, buscamos desenvolver uma distribuição bivariada juntamente com sua estrutura de dependência que possa explicar ambas as variáveis.

Assim, modelar simultaneamente duas variáveis resposta através de cópulas oferece uma abordagem computacionalmente viável para obter um modelo multivariado no contexto de regressão, permitindo a utilização de diferentes estruturas de dependência para determinar qual se ajusta melhor aos dados.

É importante ressaltar que o modelo apresenta duas variáveis resposta com distribuições marginais Beta, pertencentes à família exponencial. Conseqüentemente, precisamos estimar as médias e os parâmetros de precisão associados às distribuições para, posteriormente, obter o parâmetro de dependência. O ambiente R oferece diversas ferramentas para modelar esse tipo de distribuição com resposta limitada.

Diversas classes de cópulas podem ser utilizadas, como as elípticas e arquimedianas, conforme mostrado nas [Tabela 1](#) e [Tabela 2](#), que apresentam diferentes estruturas de dependência, bem como valores de concordância  $\tau$ . Simultaneamente é possível modelar uma dependência negativa ou positiva por meio de graus de rotação.

## 5.2 Metodologia

As cópulas são modelos convenientes, computacionalmente flexíveis e eficientes, implementadas no pacote GAMLSS (Generalized Additive Models for Location, Scale, and Shape) e utilizadas na função `copulaReg()` no pacote R `SemiParBIVProbit`. Além disso, os parâmetros são determinados em duas etapas, primeiro para as distribuições marginais e depois para a função cópula.

Seja a função de distribuição acumulada de duas variáveis aleatórias contínuas  $Z_1$  e  $Z_2$ , obtidas da análise de componentes principais e com valores entre  $(0, 1)$ . Consideremos também as covariáveis Saúde, Educação 1, Educação 2, Educação 3 e Renda obtidas de [ONPE \(2021\)](#). Seguidamente, o modelo é denotado de forma geral como uma função de distribuição acumulada:

$$F(z_1, z_2 | \mathbf{v}) = C(F_{Z_1}(z_1 | \mu_1, \phi_1), F_{Z_2}(z_2 | \mu_2, \phi_2); \zeta, \theta)$$

em que  $Z_1$  e  $Z_2$  são variáveis resposta das eleições de Peru (preferência pelo partido progressista e preferência pelo partido tradicional),  $F_{Z_1}(z_1 | \mu_1, \phi_1)$  e  $F_{Z_2}(z_2 | \mu_2, \phi_2)$  são as distribuições marginais das variáveis  $Z_1$  e  $Z_2$ ,  $\mathbf{v}$  é o vetor de parâmetros associados às distribuições marginais das variáveis resposta  $\mu_1, \phi_1, \mu_2, \phi_2$  conjuntamente com os parâmetros associados à cópula  $\zeta, \theta$ .  $C$  é definida como uma função cópula com coeficiente de dependência  $\theta$ ,  $\zeta$  representa o número de graus de liberdade da cópula t-Student (unicamente aparece em  $C$  e  $\mathbf{v}$  quando a cópula é empregada), e os parâmetros em  $\mathbf{v}$  estão conectados com as covariáveis através de preditores aditivos. Cópulas permitem duas associações de parâmetros em que  $\zeta$  pode representar um coeficiente de dependência adicional.

Os parâmetros do modelo estão relacionados às covariáveis e coeficientes de regressão através de preditores aditivos  $\eta$ 's, também conhecidos como funções de ligação, que garantem a manutenção das restrições dos espaços paramétricos. As equações das variáveis resposta são escritas como:

$$\eta_{1i} = \mathbf{U}_i^\top \boldsymbol{\beta}_1 = \beta_{01} + \beta_{11} \text{Educação}1_i + \beta_{21} \text{Renda}_i$$

$$\eta_{2i} = \mathbf{U}_i^\top \boldsymbol{\beta}_2 = \beta_{02} + \beta_{12} \text{Educação}1_i$$

em que  $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \beta_{21})$  e  $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12})$  representam as estimativas dos parâmetros associados às variáveis resposta, e cada uma mostra as covariáveis que tiveram mais influência no modelo univariado.

A estimação dos parâmetros para o modelo de cópula com marginais contínuas beta é realizada através da função de log-verossimilhança, que pode ser escrita como:

$$l(\delta) = \sum_{i=1}^n \log c(F_{Z_1}(z_{1i}|\mu_{1i}, \phi_1), F_{Z_2}(z_{2i}|\mu_{2i}, \phi_2); \theta) + \sum_{i=1}^n [\log f_{Z_1}(z_{1i}|\mu_{1i}, \phi_1) + \log f_{Z_2}(z_{2i}|\mu_{2i}, \phi_2)]$$

em que  $c$  é a densidade da cópula (os parâmetros foram suprimidos por simplicidade), dado por  $\frac{\partial^2 C(F_{Z_1}(z_{1i}), F_{Z_2}(z_{2i}))}{\partial F_{Z_1}(z_{1i}) \partial F_{Z_2}(z_{2i})}$ . Os parâmetros de distribuição são definidos como:  $\mu_{1i} = g_{\mu_1}^{-1}(\eta_{1i})$ ,  $\mu_{2i} = g_{\mu_2}^{-1}(\eta_{2i})$ ,  $\phi_1 = g_{\phi_1}^{-1}(\eta_{\phi_1})$ ,  $\phi_2 = g_{\phi_2}^{-1}(\eta_{\phi_2})$ ,  $\theta = g_{\theta}^{-1}(\eta_{\theta})$ , em que as funções  $g$  representam as funções de ligação. Além disso, o parâmetro  $\delta$  está associado aos parâmetros  $\beta_{\mu_1}^\top, \beta_{\mu_2}^\top, \beta_{\phi_1}, \beta_{\phi_2}, \beta_{\theta}$ , os quais, por sua vez, estão relacionados aos parâmetros  $\eta_{\mu_{1i}}, \eta_{\mu_{2i}}, \eta_{\phi_1}, \eta_{\phi_2}$ , e  $\eta_{\theta}$ . Seguidamente, dada a flexibilidade das estruturas dos preditores, o algoritmo de otimização é através da log-verossimilhança penalizada  $l_p(\delta)$  (MARRA; RADICE, 2017).

Estimamos os parâmetros das distribuições marginais e a cópula, considerando equações associadas as médias, duas aos parâmetros de precisão e uma equação para o parâmetro de dependência. Para elaborar cada equação, consideramos os indicadores de desenvolvimento humano de Peru (Saúde, Educação 1, Educação 2, Educação 3 e Renda) (ONPE, 2021) que tiveram maior influência em cada variável resposta do modelo univariado.

Para obter as estimativas, utilizamos a função `copulaReg()` do pacote R `SemiParBIV-Probit` para modelar as seguintes cópulas elípticas e arquimedianas: cópula Gaussiana "N", Frank "F", Ali-Mikhail-Haq "AMH", Farlie-Gumbel-Morgenstern "FGM", e outras cópulas com rotações para capturar diferentes tipos de dependência como, Gumbel com grau de rotação 90 "G90", Joe com grau de rotação 90 "J90", em que as rotações podem apresentar diferentes graus seja "0", "90", "180" e "270" e permitem modelar dependências positivas ou negativas nas caudas.

Os códigos da metodologia proposta foram desenvolvidos no software R e estão disponíveis na [Código-fonte 3](#). Além disso, podem ser acessado pelo seguinte link: [Bivariate regression via copulas](#).

### 5.3 Resultados

Conseqüentemente, foram desenvolvidos seis tipos de cópulas, cada um com sua estrutura de dependência correspondente. Na Tabela 25, empregamos os critérios AIC e BIC para a seleção do modelo que melhor descreve os dados, em que o melhor apresenta o menor valor em tais critérios. Dessa forma, a cópula Joe com rotação de 90° indica o menor valor de AIC com -38.57 e o menor valor de BIC com -28.82.

Tabela 25 – Critério de seleção de modelos

| Cópula                    | AIC     | BIC     |
|---------------------------|---------|---------|
| Normal                    | -37,375 | -27,624 |
| Frank                     | -38,476 | -28,725 |
| Ali-Mikhail-Haq           | -37,455 | -27,704 |
| Farlie-Gumbel-Morgenstern | -37,944 | -28,193 |
| 90° Gumbel                | -38,300 | -28,549 |
| 90° Joe                   | -38,568 | -28,817 |

Posteriormente, apresentam-se os resultados das estimativas para a cópula Joe 90°. Na Tabela 26, é possível observar que as variáveis Educação 1 e Renda demonstram significância, exibindo p-valores de 0,003 e 0,000, respectivamente. Além disso, os coeficientes associados revelam valores próximos aos estimados no modelo univariado, onde Educação 1 exerce um impacto positivo, enquanto Renda impacta negativamente na preferência pelos partidos progressistas.

Tabela 26 – Estimativas dos parâmetros para preferência pelos partidos progressistas

| Cópula 90° Joe           | Estimate | Std. Error | z value | Pr(> z ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | -1,163   | 1,494      | -0,778  | 0,436    |     |
| Educação 1               | 7,497    | 2,577      | 2,910   | 0,003    | **  |
| Renda                    | -12,026  | 2,212      | -5,436  | 0,000    | *** |
| logito( $\hat{\phi}_1$ ) | -1,076   | 0,261      | -4,112  | 0,000    | *** |

Nível de significância:  $p \leq 0,001$  \*\*\*;  $0,001 < p \leq 0,01$  \*\*;  $0,01 < p \leq 0,05$  \*

Na Tabela 27, destaca-se a significância da variável Educação 1, cujo p-valor é praticamente zero. O coeficiente estimado demonstra proximidade com o valor obtido no modelo univariado, indicando que Educação 1 exerce um impacto negativo na preferência pelos partidos tradicionais.

Tabela 27 – Estimativas dos parâmetros para preferência pelos partidos tradicionais

| Cópula 90° Joe           | Estimate | Std. Error | z value | Pr(> z ) |     |
|--------------------------|----------|------------|---------|----------|-----|
| Intercept.               | 8,493    | 1,365      | 6,220   | 0,000    | *** |
| Educação 1               | -13,343  | 2,103      | -6,343  | 0,000    | *** |
| logito( $\hat{\phi}_2$ ) | -1,203   | 0,269      | -4,459  | 0,000    | *** |

Nível de significância:  $p \leq 0,001$  \*\*\*;  $0,001 < p \leq 0,01$  \*\*;  $0,01 < p \leq 0,05$  \*

Na Tabela 28, são apresentadas as estimativas para o parâmetro de dependência da cópula Joe com uma rotação de 90°. Conforme evidenciado na Tabela 2, a cópula Joe exibe um parâmetro de dependência com valores positivos no intervalo  $(1, \infty)$ . Entretanto, ao ser rotacionada em 90°, esses valores extrapolam esse intervalo. A estimativa obtida, portanto, é -0,559, sugerindo uma dependência negativa moderada. Isso indica que valores elevados na preferência pelos partidos progressistas estão associados a menores preferências pelos partidos tradicionais, e vice-versa.

Adicionalmente, o valor de  $\tau$  é calculado por meio de uma transformação, conforme mostrado na Tabela 2, também exibindo um valor negativo de -0,243.

Tabela 28 – Estimativa de  $\theta$ 

| Cópula 90° Joe Intercept. | Estimate | Std. Error | t value | Pr(> t ) | $\tau$ |
|---------------------------|----------|------------|---------|----------|--------|
|                           | -0,559   | 0,756      | -0,739  | 0,460    | -0.243 |

## 5.4 Discussão final

Em resumo, após a aplicação do método de redução de dimensões nos dados eleitorais do Peru, observamos uma distribuição particular das regiões. Duas componentes emergiram, explicando a preferência pelos partidos progressistas e pelos partidos tradicionais, respectivamente. Posteriormente, incorporamos variáveis explicativas dos indicadores de desenvolvimento humano e dada a presença dessas variáveis em cada componente, aplicamos um modelo de regressão bivariada via cópulas para identificar a dependência entre as componentes.

As cópulas representam uma ferramenta estatística versátil, proporcionando flexibilidade computacional. Neste contexto, possibilitam a construção de uma distribuição bivariada para duas variáveis dependentes limitadas. Diversos pacotes em R foram empregados para realizar a análise e adaptar as respostas limitadas no modelo. Além disso, diferentes tipos de cópulas, como cópulas elípticas e arquimedianas, foram explorados, permitindo a identificação da cópula que melhor se ajusta aos dados.

Com base nos critérios AIC e BIC, concluímos que a cópula arquimediana denominada Joe, com rotação de 90°, apresenta o melhor ajuste aos dados. Como resultado, as estimativas para os parâmetros associados a cada variável explicativa são consistentes com os modelos univariados previamente apresentados, sendo também estatisticamente significativos. Além disso, obtivemos uma estimativa para o parâmetro de dependência, que possui um valor de -0,559, indicando uma dependência negativa entre a preferência pelos partidos progressistas e a preferência pelos partidos tradicionais na presença das variáveis explicativas.

---

## CONSIDERAÇÕES FINAIS

---

A metodologia proposta abrange oito etapas e foi aplicada nas eleições presidenciais do Peru em 2021, utilizando dados apresentados pela [ONPE \(2021\)](#), e nas eleições do Brasil em 2022, com informações fornecidas pelo Tribunal Superior Eleitoral [TSE \(2022\)](#).

Inicialmente, os dados foram transformados em dados composicionais, caracterizados por valores proporcionais e uma soma constante. No entanto, esses dados possuem restrições adicionais, pois estão em um espaço limitado conhecido como Simplex. Essas restrições tornam a análise estatística mais desafiadora, requerendo métodos de transformação adequados. Nesse cenário, aplicamos o método proposto por [Aitchison \(1982\)](#), que envolve uma transformação de razões logarítmicas para eliminar a dependência das variáveis, superando as restrições mencionadas. Devido à dimensionalidade dos dados, optamos por realizar uma análise de componentes principais para reduzir as dimensões, embora outras técnicas de redução de dimensionalidade, como a análise de fatores (AF), também podem ser consideradas. Após a obtenção dos escores, notamos uma variabilidade diferente em cada componente. Como resultado, realizamos uma transformação para valores no intervalo unitário.

Posteriormente, foram coletadas variáveis explicativas relacionadas a fatores sociais na decisão de votos. No entanto, é viável ajustar esta etapa incorporando variáveis explicativas adicionais, como indicadores socioeconômicos, políticos ou ambientais, dependendo da circunscrição estudada.

Em seguida, implementamos um modelo de regressão Beta [Ferrari e Cribari-Neto \(2004\)](#). No entanto, vale ressaltar a possibilidade de explorar outros modelos de regressão com resposta limitada, onde a variável resposta segue uma distribuição Simplex ([BARNDORFF-NIELSEN; JØRGENSEN, 1991](#)), [Kumaraswamy \(KUMARASWAMY, 1980\)](#) e [Johnson \(JOHNSON, 1949\)](#), visando identificar o que melhor se adequa aos dados disponíveis.

Da mesma forma, no diagnóstico de resíduos, utilizamos a análise de resíduos quantílicos normalizados ([DUNN; SMYTH, 1996](#)). No entanto, é possível empregar outras técnicas

adicionais de diagnóstico para analisar, identificar pontos influentes e detectar eventuais desvios do modelo, como leverage generalizado (WEI *et al.*, 1998) e discrepância de ajuste (FERRARI; CRIBARI-NETO, 2004).

Além disso, desenvolvemos um modelo de regressão bivariada adotando a abordagem das cópulas (NELSEN, 2006) para os dados das eleições no Peru em 2021. O modelo considera a correlação entre as variáveis de resposta (preferência por partidos progressistas e tradicionais) na presença de variáveis explicativas. Embora a correlação entre as variáveis de resposta não é significativa, o modelo evidencia uma relação entre elas. Assim também, as variáveis foram obtidas por meio da análise de componentes principais, e conseqüentemente elas são linearmente independentes e ortogonais no espaço. No entanto, essa independência linear e ortogonal não implica independência total, uma vez que as regiões analisadas estão próximas umas das outras e compartilham as mesmas variáveis explicativas.

Portanto, em futuros trabalhos, é possível desenvolver abordagens alternativas para analisar a correlação entre as regiões (PRATES *et al.*, 2022) a partir de uma perspectiva espaço-temporal, utilização de campos aleatórios de Markov (PRATES *et al.*, 2015) para lidar com estruturas de dependência e a incorporação de efeitos aleatórios (PRATES *et al.*, 2013).

Adicionalmente, essa metodologia pode ser aplicada em diversas circunscrições, como distritos, regiões ou estados, permitindo a compreensão do comportamento político em diferentes sociedades. Isso possibilita analisar quais covariáveis determinam a decisão da população ao escolher votar em um candidato presidencial. A análise pode ser conduzida tanto sob uma abordagem clássica quanto sob uma abordagem bayesiana.

## REFERÊNCIAS

---

- AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 44, n. 2, p. 139–160, 1982. Citado nas páginas 15, 19 e 72.
- AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 38.
- ANJOS, U. U. d.; FERREIRA, F. H.; KOLEV, N.; MENDES, B. V. d. M. Modelando dependências via cópulas. 2004. Citado nas páginas 33 e 37.
- ATLASBR. **Atlas do Desenvolvimento Humano no Brasil**. 2022. Disponível em: <<http://www.atlasbrasil.org.br/ranking>>. Acesso em: 20 Jun 2022. Citado na página 63.
- BARNDORFF-NIELSEN, O. E.; JØRGENSEN, B. Some parametric models on the simplex. **Journal of multivariate analysis**, Elsevier, v. 39, n. 1, p. 106–116, 1991. Citado na página 72.
- BAYER, F. M. Modelagem e inferência em regressão beta. Universidade Federal de Pernambuco, 2011. Citado na página 25.
- BAZAN, J. L.; SULMONT, D.; CALDERÓN, A. Las organizaciones políticas en las elecciones presidenciales peruanas de 2011 usando análisis de componentes principales. **Revista de Estudios Sociales**, v. 14, n. 27, p. 10–27, 2012. Citado nas páginas 15 e 16.
- BIONDO, T. R.; SUZUKI, A. K. Modelos de sobrevivencia bivariados derivados da copula arquimediana de clay-ton: Uma abordagem bayesiana. **Matemática e Estatística em Foco**, v. 4, n. 2, p. 87–102, 2016. Citado na página 32.
- BOOGAART, K. G. Van den; TOLOSANA-DELGADO, R. “compositions”: a unified r package to analyze compositional data. **Computers & Geosciences**, Elsevier, v. 34, n. 4, p. 320–338, 2008. Citado nas páginas 43 e 60.
- BOUYÉ, E.; DURRLEMAN, V.; NIKEGHBALI, A.; RIBOULET, G.; RONCALLI, T. Copulas for finance-a reading guide and some applications. **Available at SSRN 1032533**, 2000. Citado na página 32.
- CHERUBINI, U.; LUCIANO, E.; VECCHIATO, W. **Copula methods in finance**. West Sussex, England: John Wiley & Sons, 2004. Citado nas páginas 34 e 35.
- CZADO, C. Analyzing dependent data with vine copulas. **Lecture Notes in Statistics, Springer**, Springer, v. 222, 2019. Citado na página 35.
- DANAHER, P. J.; SMITH, M. S. Modeling multivariate distributions using copulas: Applications in marketing. **Marketing science**, INFORMS, v. 30, n. 1, p. 4–21, 2011. Citado na página 32.
- DENUIT, M.; PURCARU, O.; KEILEGOM, I. V. Bivariate archimedean copula models for censored data in non-life insurance. 2006. Citado na página 32.

- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and graphical statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado nas páginas 28, 53 e 72.
- EGOZCUE, J. J.; PAWLOWSKY-GLAHN, V.; MATEU-FIGUERAS, G.; BARCELO-VIDAL, C. Isometric logratio transformations for compositional data analysis. **Mathematical geology**, Springer, v. 35, n. 3, p. 279–300, 2003. Citado nas páginas 19 e 20.
- EMBRECHTS, P.; LINDSKOG, F.; MCNEIL, A. Modelling dependence with copulas. **Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich**, v. 14, p. 1–50, 2001. Citado na página 32.
- FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado nas páginas 23, 28, 72 e 73.
- FISHER, R. A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character**, The Royal Society, v. 222, p. 309–368, 1922. Citado na página 24.
- GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, Oxford University Press, v. 58, n. 3, p. 453–467, 1971. Citado nas páginas 44 e 60.
- GODA, K.; TESHAMARIAM, S. **Multi-variate seismic demand modelling using copulas: Application to non-ductile reinforced concrete frame in Victoria, Canada**. Amsterdam, Netherlands: Elsevier, 2015. 39–51 p. Citado na página 32.
- INEI. **ÍNDICE DE DESARROLLO HUMANO – IDH**. 2019. Disponível em: <<https://www.ipe.org.pe/portal/indice-de-desarrollo-humano-idh/comment-page-9/>>. Acesso em: 20 Jun 2022. Citado na página 50.
- ISO3166. **ISO 3166-2 codes**. 1998. Disponível em: <<https://www.iso.org/obp/ui/#iso:code:3166:PE>>. Acesso em: 24 Nov 2020. Citado nas páginas 18 e 45.
- JOE, H. **Dependence modeling with copulas**. FL, USA: CRC press, 2014. Citado na página 32.
- JOHNSON, N. L. Systems of frequency curves generated by methods of translation. **Biometrika**, JSTOR, v. 36, n. 1/2, p. 149–176, 1949. Citado na página 72.
- JOHNSON, R. A.; WICHERN, D. W. *et al.* Applied multivariate statistical analysis. Prentice hall Upper Saddle River, NJ, 2002. Citado na página 20.
- KUMARASWAMY, P. A generalized probability density function for double-bounded random processes. **Journal of hydrology**, Elsevier, v. 46, n. 1-2, p. 79–88, 1980. Citado na página 72.
- MARRA, G.; RADICE, R. Bivariate copula additive models for location, scale and shape. **Computational Statistics & Data Analysis**, Elsevier, v. 112, p. 99–113, 2017. Citado na página 69.
- MOSIMANN, J. E. On the compound multinomial distribution, the multivariate  $\beta$ -distribution, and correlations among proportions. **Biometrika**, JSTOR, v. 49, n. 1/2, p. 65–82, 1962. Citado na página 15.

- NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society Series A: Statistics in Society**, Oxford University Press, v. 135, n. 3, p. 370–384, 1972. Citado na página 27.
- NELSEN, R. B. **An introduction to copulas**. New York, USA: Springer, 2006. Citado nas páginas 31, 34, 36, 38 e 73.
- OLIVEIRA, M. S. de. **Um Modelo de Regressão Beta: teoria e aplicações**. Tese (Doutorado) — Instituto de Matemática e Estatística da Universidade de São Paulo, 12/04/2004., 2004. Citado na página 26.
- ONPE. **PRESENTACIÓN DE RESULTADOS Elecciones Generales y Parlamento Andino 2021**. 2021. Disponível em: <<https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T>>. Acesso em: 20 Jun 2022. Citado nas páginas 39, 68, 69 e 72.
- PALEOLOGOU, S.-M. Income and democracy: a bivariate copula approach. **Journal of Applied Statistics**, Taylor & Francis, v. 50, n. 7, p. 1635–1649, 2023. Citado na página 32.
- PEARSON, K. On lines and planes of closest fit to system of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 2, p. 559–572, 1901. Citado na página 20.
- PRATES, M. O.; AZEVEDO, D. R.; MACNAB, Y. C.; WILLIG, M. R. Non-separable spatio-temporal models via transformed multivariate gaussian markov random fields. **Journal of the Royal Statistical Society Series C: Applied Statistics**, Oxford University Press, v. 71, n. 5, p. 1116–1136, 2022. Citado na página 73.
- PRATES, M. O.; DEY, D. K.; WILLIG, M. R.; YAN, J. Transformed gaussian markov random fields and spatial modeling of species abundance. **Spatial Statistics**, Elsevier, v. 14, p. 382–399, 2015. Citado nas páginas 32 e 73.
- PRATES, M. O.; JR, R. H. A.; DEY, D. K.; YAN, J. Assessing intervention efficacy on high-risk drinkers using generalized linear mixed models with a new class of link functions. **Biometrical Journal**, Wiley Online Library, v. 55, n. 6, p. 912–924, 2013. Citado na página 73.
- RODRIGUES, P. C.; LIMA, A. T. Analysis of an european union election using principal component analysis. **Statistical Papers**, Springer, v. 50, p. 895–904, 2009. Citado nas páginas 16, 21 e 22.
- SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978. Citado na página 38.
- SEN, P. K.; SINGER, J. M. **Large Sample Methods in Statistics: An Introduction with Applications**. [S.l.]: Chapman & Hall, 1993. Citado na página 26.
- SKLAR, A. Fonctions de répartition à plusieurs dimensions et leurs marges. **Publications de l’Institut Statistique de l’Université de Paris**, v. 8, p. 229–239, 1959. Citado nas páginas 31 e 32.
- STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. D. **Flexible regression and smoothing: using GAMLSS in R**. Florida, United States: CRC Press, 2017. Citado na página 50.

TSE. **Eleição Geral Ordinária**. 2022. Disponível em: <<https://www.tse.jus.br/comunicacao/noticias/2022/Outubro/100-das-secoes-totalizadas-confira-como-ficou-o-quadro-eleitoral-apos-o-1o-turno>> Acesso em: 20 Jun 2022. Citado nas páginas 56 e 72.

WATTS, S. The gaussian copula and the financial crisis: A recipe for disaster or cooking the books? **University of Oxford**, v. 8, 2016. Citado na página 35.

WEI, B.-C.; SHI, J.-Q.; FUNG, W.-K.; HU, Y.-Q. Testing for varying dispersion in exponential family nonlinear models. **Annals of the Institute of Statistical Mathematics**, Springer, v. 50, p. 277–294, 1998. Citado nas páginas 28 e 73.

---

## ***CÓDIGOS R***

---

---

### **Código-fonte 1** – Código do modelo em R das eleições peruanas

---

```
library(compositions)
library(Hmisc, pos=4)
library(foreign, pos=4)
library(corrplot)
library(ggplot2)
library(readxl)

#1.Dados eleições 1er turno

setwd(dirname(rstudioapi::getActiveDocumentContext())$path))
Electionsdata2021 <- read_excel("elections2021_1round.xlsx")
names(Electionsdata2021)
head(Electionsdata2021)

#2.Dados eleições 2do turno
setwd(dirname(rstudioapi::getActiveDocumentContext())$path))
Elections2021_2round <- read_excel("Elections2021_2round.xlsx")
Elections2021_2round

#3. Transformações
### Transformação em proporções

Matriz_2round = as.matrix(Elections2021_2round [,c(3,4,5)])
```

```
Matriz_2round_prop =matrix(numeric(25),nrow = 25,ncol =2)
for(i in 1:2){
  Matriz_2round_prop[,i] = Matriz_2round[,i]/Matriz_2round[,3]
}
colnames(Matriz_2round_prop) = c("Party1","Party2")

data2_2021_prop = as.data.frame(Matriz_2round_prop)
data2_2021_prop

parties=c("PL","FP","RP","AVP","ACCP","OTHERS","VALID VOTES")
matriz_votes_2021 = as.matrix(Electionsdata2021[,parties])

matriz_prop_2021 =matrix(numeric(25),nrow = 25,ncol = 6)
matriz_prop_2021

for(i in 1:6){
  matriz_prop_2021[,i] = (matriz_votes_2021[,i]/matriz_votes_2021[,7])*100
}
matriz_prop_2021

### Transformação clr
parties1=c("PL","FP","RP","AVP","ACCP","OTHERS")
matriz_tr_2021 = matrix(numeric(25),nrow = 25,ncol = 6)
colnames(matriz_tr_2021)= parties1
matriz_tr_2021=clr(matriz_prop_2021)
matriz_tr_2021

# 4.Análise de componentes principais dos dados transformados

TransformedCPA<- princomp(matriz_tr_2021,cor =TRUE)
TransformedCPA$loadings
summary(TransformedCPA)
TransformedCPA$scores

biplot(TransformedCPA,xlab='First Component',
        ylab='Second Component',main ="Log contrast PCA scale1")
```

```
# 5. Normalização dos escores
```

```
#getAnywhere(biplot.princomp)
lam =((TransformedCPA$sdev))*sqrt(TransformedCPA$n.obs)
lam
#scale != 0 >> lam^scale
#scale = 1 so, we say lam = lam, then
lam = lam
score1 = TransformedCPA$scores[,1]/ lam[1]
score2 =TransformedCPA$scores[,2] / lam[2]
score1 #hist(score1, main = "Histogram Score 1",xlim=c(-0.6,0.6))
score2 #hist(score2, main = "Histogram Score 2",xlim=c(-0.6,0.6))
```

```
### escore 1
```

```
score1_t1 <- (score1 - min(score1))/(max(score1) - min(score1))
score1_t1[15]= score1_t1[15]+0.0001
score1_t1[21] =score1_t1[21] -0.0001
score1_t1 #hist(score1_t1, main = "Histograma de z1",ylim = c(0, 6),xlim = c(0, 1.0)
```

```
### escore 2
```

```
score2_t1 <- (score2 - min(score2))/(max(score2) - min(score2))
score2_t1[23]= score2_t1[23]+0.0001
score2_t1[1] =score2_t1[1] -0.0001
score2_t1 #hist(score2_t1, main = "Histograma de z2",ylim = c(0, 6),xlim = c(0, 1.0)
```

```
par(mfrow=c(1,2))
```

```
par(mar=c(2,2,2,2))
```

```
# 6. Dados para implementar o modelo
```

```
#indicators
```

```
setwd(dirname(rstudioapi::getActiveDocumentContext())$path))
```

```
indicators <- as.data.frame(read_excel("indicators.xlsx"))
```

```
indicators
```

```

indicators = indicators[,c(-1,-2)]
Data_indicators = cbind(indicators,score1_t1,score2_t1 )
Data_indicators
summary(Data_indicators)

# 7. GAMLSS

#packageVersion("gamlss")

###Escore1
#####
library(gamlss)
gamlss_score1_t1_m1<- gamlss(score1_t1~HealthINEI+educationINEI1+educationINEI2+Inc
summary(gamlss_score1_t1_m1)

gamlss_score1_t1_m2<- gamlss(score1_t1~HealthINEI+educationINEI1+IncomeINEI,family=
summary(gamlss_score1_t1_m2)

#library(xtable)
#xtable(gamlss_score1_t1_m2)
gamlss_score1_t1_m3<- gamlss(score1_t1~educationINEI1+IncomeINEI,family=BE,data=Dat
summary(gamlss_score1_t1_m3)

###Escore2
#####
gamlss_score2_t1_m1<- gamlss(score2_t1 ~ HealthINEI+educationINEI1+educationINEI2+
summary(gamlss_score2_t1_m1)

gamlss_score2_t1_m2<- gamlss(score2_t1 ~ educationINEI1+educationINEI2+IncomeINEI,
summary(gamlss_score2_t1_m2)

gamlss_score2_t1_m3<- gamlss(score2_t1 ~ educationINEI1,family=BE,data=Data_indica
summary(gamlss_score2_t1_m3)
fitted(gamlss_score2_t1_m3)

```

## # 8. Diagnóstico

```

par(mfrow=c(2,2))
plot(fitted(gamlss_score1_t1_m3),resid(gamlss_score1_t1_m3),xlab="Fitted values",yl
plot(1:nrow(Data_indicators),resid(gamlss_score1_t1_m3),xlab="Index",ylab="Residual
plot(density(resid(gamlss_score1_t1_m3)),main="c. Density Estimate")#### graph 3
library(car)
qqPlot(resid(gamlss_score1_t1_m3), ylab = "Normalized quantile residuals", main ="d

par(mfrow=c(2,2))
plot(fitted(gamlss_score2_t1_m2),residuals(gamlss_score2_t1_m2),xlab="Fitted values
plot(1:nrow(Data_indicators),resid(gamlss_score2_t1_m2),xlab="Index",ylab="Residual
plot(density(resid(gamlss_score2_t1_m2)),main="c. Density Estimate")#### graph 3
qqPlot(resid(gamlss_score2_t1_m2), ylab = "Normalized quantile residuals", main ="d

```

---

**Código-fonte 2** – Código do modelo em R das eleições brasileiras

---

```

library(compositions)
library(Hmisc, pos=4)
library(foreign, pos=4)
library(corrplot)
library(ggplot2)
library(readxl)

#1.Dados eleições 1er turno no Brasil

setwd(dirname(rstudioapi::getActiveDocumentContext())$path))
ElectionsdataBrazil <- read_excel("votes2022Brazil.xlsx")
#View(ElectionsdataBrazil)
ElectionsdataBrazil
UF=ElectionsdataBrazil[,1]
State=ElectionsdataBrazil[,2]
#sum(ElectionsdataBrazil[,6])/sum(ElectionsdataBrazil[,10])
#View(ElectionsdataBrazil)

```

```

ElectionsdataBrazil <-ElectionsdataBrazil[,c(-1,-2)]
dim(ElectionsdataBrazil)
Others=as.matrix(ElectionsdataBrazil[,8]) -
  (as.matrix(ElectionsdataBrazil[,1]) +as.matrix(ElectionsdataBrazil[,2]) +
  as.matrix(ElectionsdataBrazil[,3]) +as.matrix(ElectionsdataBrazil[,4]) )

databrazil<- cbind(as.matrix(ElectionsdataBrazil[,c(1,2,3,4)]),Others
, ElectionsdataBrazil[,8])
colnames(databrazil)= c("PT","PL","MDB","PDT","OTHERS","TOTAL")
databrazil
#matriz_prop_br =matrix(numeric(27),nrow = 27,ncol = 7)
#matriz_prop_br
col_prop1<-(databrazil[,1]/databrazil[,6])*100
col_prop2<-(databrazil[,2]/databrazil[,6])*100
col_prop3<-(databrazil[,3]/databrazil[,6])*100
col_prop4<-(databrazil[,4]/databrazil[,6])*100
col_prop5<-(databrazil[,5]/databrazil[,6])*100

dados1<-cbind(UF,State,databrazil)

#2. Transformações
### Transformação em proporções
matriz_prop_br<-cbind(col_prop1,col_prop2,col_prop3,col_prop4,col_prop5)
matriz_prop_br
colnames(matriz_prop_br)= c("PT","PL","MDB","PDT","OTHERS")
matriz_prop_br

matriz_tr_br=clr(matriz_prop_br)
matriz_tr_br

rownames(matriz_tr_br)<-c("AC","AL","AP","AM","BA","CE","DF","ES",
"GO","MA","MT","MS","MG","PA","PB","PR","PE","PI",
"RJ","RN","RS","RO","RR","SC","SP","SE","TO")

### Transformação clr
TransformedCPA_br<- princomp(matriz_tr_br,cor =TRUE)
summary(TransformedCPA_br)
cargas=TransformedCPA_br$loadings
TransformedCPA_br$scores

```

```
par(mfrow = c(1,2))
biplot(TransformedCPA_br,xlab='First Component',
ylab='Second Component',main ="Biplot of the Brazilian
elections 2022",cex = 0.5)

# 4. Normalização dos escores

lam =((TransformedCPA_br$sdev))*sqrt(TransformedCPA_br$n.obs)
lam
#scale != 0 >> lam^scale
lam = lam
score1_br = TransformedCPA_br$scores[,1]/ lam[1]
score2_br =TransformedCPA_br$scores[,2] / lam[2]
score1_br
score2_br
#####standardizing scores
score1_t1 <- (score1_br - min(score1_br))/(max(score1_br) - min(score1_br))
score1_t1
score1_t1[23]= score1_t1[23]+0.0001
score1_t1[6] =score1_t1[6] -0.0001
score1_t1#hist(score1_t1, main = "Histograma de z1",
ylim = c(0, 8), xlim = c(0, 1.0),breaks = 10,
xlab = "", ylab = "")

score2_t1 <- (score2_br - min(score2_br))/(max(score2_br) - min(score2_br))
score2_t1
score2_t1[6]= score2_t1[6]+0.0001
score2_t1[17] =score2_t1[17] -0.0001
score2_t1#hist(score2_t1, main = "Histograma de z2"
,ylim = c(0, 6),xlim = c(0, 1.0),breaks = 10)

#5. Dados para implementar o modelo
```

```

setwd(dirname(rstudioapi::getActiveDocumentContext())$path)
indicators <- as.data.frame(read_excel("IDHbrasil.xlsx"))
UF = indicators[,1]
str(indicators)
summary(indicators)
indicators<- indicators[,c(3,4,5)]
indicators<-cbind(as.numeric(indicators[,1]),as.numeric(indicators[,2]),
as.numeric(indicators[,3]))

colnames(indicators)= c("renda","educação","longevidade")
renda = cbind(indicators[,1])
educação = cbind(indicators[,2])
longevidade = cbind(indicators[,3])

Data_indicators =as.data.frame(cbind(indicators,score1_t1,score2_t1 ))
summary(Data_indicators)

# 6. GAMLSS
library(gamlss)
###Escore1
#####
score1_m1<- gamlss(score1_t1 ~ renda+educação+
longevidade, family= BE,data=Data_indicators )
summary(score1_m1)

score1_m2<- gamlss(score1_t1~ renda+educação,family=BE,data=Data_indicators )
summary(score1_m2)

# 7. Diagnóstico
par(mfrow=c(2,2))
plot(fitted(score1_m2),residuals(score1_m2),xlab="Fitted values",ylab="Residuals",
main="a. Fitted values vs Residuals",
ylim = c(-3,3))#### graph 1
plot(1:nrow(Data_indicators),resid(score1_m2),xlab="Index",
ylab="Residuals",main="b. Index vs Residuals",ylim = c(-3,3))#### graph 2
plot(density(resid(score1_m2)),main="c. Density Estimate")#### graph 3
qqPlot(resid(score1_m2), ylab = "Normalized quantile residuals")

```

```
, main = "d. Normal Q-Q Plot",
ylim = c(-3,3), id = list(method = "y", n = 25, cex = 0.5), )#### graph 4
```

---

**Código-fonte 3** – Código do modelo em R das eleições Peruanas utilizando cópulas

---

```
#####Copulas#####
library(SemiParBIVProbit)
library(copula)
library(VineCopula)

library(gamlss)
library(compositions)
library(Hmisc, pos=4)
library(foreign, pos=4)
library(corrplot)
library(ggplot2)
library(readxl)
library(psych)

newdata<- read.csv("newdata.csv")
head(newdata)
attach(newdata)#Pacotes

# MODELO 1 VIA COPULAS PARA MODELAR AS MEDIAS E OBTENER VARIANCIAS, THETA E TAU
eq.mu.1 <- z1 ~ education1+income
eq.mu.2 <- z2 ~ education1
eq.sigma2.1 <- ~1
eq.sigma2.2 <- ~1
eq.theta <- ~ 1
fl <- list(eq.mu.1, eq.mu.2, eq.sigma2.1, eq.sigma2.2, eq.theta)
#Gaussian
system.time(outN <- copulaReg(fl,
margins = c("BE", "BE"),
data = newdata, gamlssfit = TRUE))
summary(outN)

system.time(outF <- copulaReg(fl,
margins = c("BE", "BE"), BivD = "F", data = newdata, gamlssfit = TRUE))
```

```
summary(outF)#sample size small
```

```
system.time(outAMH <- copulaReg(fl,  
margins = c("BE", "BE"), BivD = "AMH",data = newdata, gamlssfit = TRUE))  
summary(outAMH)
```

```
system.time(outFGM <- copulaReg(fl,  
margins = c("BE", "BE"), BivD = "FGM",data = newdata, gamlssfit = TRUE))  
summary(outFGM)
```

```
system.time(outG90 <- copulaReg(fl,  
margins = c("BE", "BE"), BivD = "G90",data = newdata, gamlssfit = TRUE))  
summary(outG90)
```

```
system.time(outJ90 <- copulaReg(fl,  
margins = c("BE", "BE"), BivD = "J90",data = newdata, gamlssfit = TRUE))  
summary(outJ90)
```

```
#####summary AIC/BIC
```

```
aic =AIC(outN,outF,outAMH,outFGM,outG90,outJ90)  
bic = BIC(outN,outF,outAMH,outFGM,outG90,outJ90)
```

