UNIVERSIDADE
EDUARDO
MONDLANE

FACULDADE DE CIÊNCIAS

Departamento de Matemática e Informática

Trabalho de Licenciatura em

Estatística

Análise dos factores associados à ocorrência do Desemprego na Cidade e Província de Maputo

Autor: Kenny Manuel Mutar

Maputo, Julho de 2025

UNIVERSIDADE
EDUARDO
MONDLANE

FACULDADE DE CIÊNCIAS

Departamento de Matemática e Informática

Trabalho de Licenciatura em

Estatística

Análise dos factores associados à ocorrência do Desemprego na Cidade e Província de Maputo

Autor: **Kenny Manuel Mutar**

Supervisor: Bonifácio José, MSc, UEK

Maputo, Julho de 2025

Dedicatória

Dedico este trabalho aos meus pais

(**Manuel Mutar e Verónica Paulo Deroteia**)

Declaração de Honra

Declaro por minha honra que o presente Trabalho de Licenciatura é resultado da minha investigação e que o processo foi concebido para ser submetido apenas para a obtenção do grau de Licenciado em Estatística, na faculdade de Ciências da Universidade Eduardo Mondlane.

Maputo, Julho de 2025

_____

Kenny Manuel Mutar

Agradecimentos

Em primeiro lugar, agradeço a **DEUS**, por ter me dado saúde e força, para que eu chegasse até aqui.

Aos meus pais, que sempre estiveram presentes em todos os momentos da minha vida, dando apoio e segurança.

Ao Meu Irmão **Abel Geraldo Coutinho** e a minha Prima **Silvana Lia**, pelo apoio incondicional e incentivo ao longo de toda a minha jornada acadêmica.

Ao meu Supervisor, **Bonifácio José**, pela orientação, paciência e dedicação durante o desenvolvimento deste trabalho, sempre me guiando com muito profissionalismo.

Estendo meus agradecimentos aos professores do curso de Estatística, que contribuíram significativamente para minha formação acadêmica e profissional, compartilhando seus conhecimentos e experiências.

Sou grato aos meus colegas e amigos, especialmente ao **Valdemiro Américo**, **Paulo Notiço**, **Élio Confiança**, pelo apoio, companheirismo e pelas trocas de ideias que enriqueceram minha experiência acadêmica.

Ao **Programa Crescimento Inclusivo em Moçambique** por todo suporte técnico que tornou possível a conclusão deste trabalho.

A todos, meu muito obrigado.

Resumo

O desemprego é definido pela Organização Internacional do Trabalho (OIT) como a situação em que indivíduos em idade activa não estão empregados, mas estão disponíveis e procuram trabalho activamente. Este fenómeno é um dos principais desafios socioeconómicos contemporâneos, afectando negativamente a qualidade de vida e aumentando as desigualdades sociais. O objectivo deste trabalho é analisar os factores que influenciam a ocorrência do desemprego na Cidade e Província de Maputo. Para isso, foram utilizados dados secundários provenientes do Inquérito sobre Orçamento Familiar (IOF) de 2022, realizado em Moçambique, concretamente na Cidade e Província de Maputo, com uma amostra de 9761 indivíduos. Para verificar a associação entre a variável dependente e as variáveis independentes, foi aplicado o teste qui quadrado de independência. Posteriormente, foi utilizada a regressão logística binária para identificar e analisar os factores associados à ocorrência do desemprego, com o teste de Hosmer e Lemeshow indicando um bom ajuste do modelo. Os resultados deste trabalho indicam que factores como a idade, nível de escolaridade, estado civil, região de residência e posição no agregado familiar, estão associados a ocorrência do desemprego na Cidade e Província de Maputo.

Palavras-chave: Desemprego, regressão logística, mercado de trabalho.

Abstract

Unemployment is defined by the International Labour Organization (ILO) as the situation in which individuals of working age are not employed but are available and actively seeking work. This phenomenon is one of the main contemporary socioeconomic challenges, negatively affecting quality of life and increasing social inequalities. The objective of this work is to analyze the factors that influence the occurrence of unemployment in the City and Province of Maputo. For this purpose, secondary data from the 2022 Household Budget Survey (IOF), carried out in Mozambique, specifically in the City and Province of Maputo, with a sample of 9761 individuals, were used. To verify the association between the dependent variable and the independent variables, the chi-square test of independence was applied. Subsequently, binary logistic regression was used to identify and analyze the factors associated with the occurrence of unemployment, with the Hosmer and Lemeshow test indicating a good fit of the model. The results of this work indicate that factors such as age, level of education, marital status, region of residence and position in the household are associated with the occurrence of unemployment in the City and Province of Maputo.

Keywords: Unemployment, logistic regression, labor market

Lista de Abreviaturas

PEA	População Economicamente Activa
OIT	Organização Internacional do Trabalho
INE	Instituto Nacional de Estatística
IDA	Associação Internacional de Desenvolvimento
MGCAS	Ministério do Género, Criança e Acção Social
FEC	Fundação Fé e Cooperação
MLG	Modelos Lineares Generalizados
TRV	Teste da Razão de Verossimilhança
IOF	Inquérito Sobre Orçamento Familiar
UPA	Unidades Primárias de Amostragem
AE	Área de Enumeração
OR	Razão de chances
AUC	Área Sob a Curva

Índice

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

INTRODUÇÃO

1.1 Contextualização

A população economicamente activa (PEA) é fundamental para o mercado de trabalho, englobando pessoas em idade laboral que estão empregadas ou em busca de emprego. Nos últimos anos, a PEA global enfrentou desafios significativos, como mudanças tecnológicas e crises económicas, que redefiniram a natureza do trabalho. A Organização Internacional do Trabalho (OIT, 2020) destaca que essas transformações impactaram as oportunidades de emprego e as condições de trabalho em escala global.

O desemprego é um dos principais problemas socioeconómicos contemporâneos, afectando desigualmente diferentes regiões e grupos demográficos. Em 2020, a taxa de desemprego global atingiu 6,5%, representando cerca de 220 milhões de pessoas desempregadas, em grande parte devido à pandemia de COVID-19 (OIT, 2021). Este cenário evidencia a vulnerabilidade do mercado de trabalho a choques externos e como as desigualdades existentes podem ser exacerbadas em tempos de crise.

Na África, o desemprego apresenta um panorama desafiador, especialmente entre os jovens. O Fórum Económico Mundial (2022) destaca que mais de 60% da população africana está abaixo dos 25 anos, o que intensifica a pressão sobre o mercado de trabalho e as oportunidades de emprego no continente. Apesar do crescimento económico, a criação de empregos não acompanha o aumento da força de trabalho. Segundo dados do Banco Mundial (2023), a taxa de desemprego juvenil na África Subsaariana é de 10,2%.

Moçambique, inserido neste contexto africano mais amplo, enfrenta seus próprios desafios únicos em relação ao emprego. Com uma população estimada em 30,8 milhões de habitantes em 2021, dos quais dois terços têm menos de 25 anos de idade (United Nations Mozambique, 2021), o país lida com uma pressão crescente para criar oportunidades de emprego para sua população jovem em expansão.

A análise dos factores associados ao desemprego é crucial para o desenvolvimento de políticas públicas eficazes e estratégias de criação de emprego. Compreender as dinâmicas específicas do mercado de trabalho local pode contribuir para intervenções mais direccionadas e eficientes.

1.2 Definição do Problema

O desemprego é um desafio socioeconómico crítico em todo o mundo, responsável por altas taxas de pobreza e desigualdade social. O peso deste problema está se intensificando em muitas regiões, mesmo à medida que alguns países experimentam crescimento económico.

O elevado número de pessoas desempregadas, constitui uma grande preocupação em Moçambique. Por isso, ao longo dos anos, tem havido planos estratégicos desenvolvidos pelo Governo de Moçambique e organizações internacionais para minimizar esse problema no país.

De acordo com o Instituto Nacional de Estatística (INE,2023), a província de Maputo e a cidade de Maputo apresentam taxas de desemprego de 33,1% e 36,5%, respectivamente, significativamente acima da média nacional.

A concentração de altas taxas de desemprego na área metropolitana de Maputo é particularmente preocupante, considerando que esta região é o centro económico e político do país. Este cenário revela uma crise laboral aguda, onde uma parcela substancial da população em idade activa se encontra sem ocupação formal, impactando directamente a qualidade de vida e o desenvolvimento socioeconómico da região.

Daí que surge a seguinte pergunta de pesquisa: "Quais são os factores que influenciam na ocorrência do desemprego na Cidade e Província de Maputo?"

1.3 Objectivos

1.3.1 Objectivo Geral

Analisar os factores que influenciam na ocorrência do desemprego na Cidade e Província de Maputo

1.3.2 Objectivos Específicos

- Descrever o perfil sociodemográfico dos indivíduos com e sem emprego;
- Verificar a associação bivariada entre a variável de interesse e as demais variáveis em estudo;

- Estimar um modelo probabilístico que melhor relaciona o desemprego e os potenciais factores.
- Identificar os factores associados à ocorrência do desemprego.

1.4 Relevância do Estudo

A relevância deste estudo sobre os factores associados ao desemprego na cidade e província de Maputo fundamenta-se em diversos aspectos interligados. Considerando a posição de Maputo como centro económico e político de Moçambique, as questões relacionadas ao emprego nesta região têm repercussões que se estendem por todo o território nacional. Este estudo preenche uma lacuna significativa na literatura existente sobre o mercado de trabalho moçambicano, oferecendo análises actualizadas e específicas. Os resultados obtidos têm potencial para fundamentar a elaboração de políticas públicas mais eficientes na geração de emprego e redução da pobreza, não apenas em Maputo, mas em todo o país. As conclusões deste estudo podem orientar organizações e empresas locais no alinhamento de suas estratégias com as necessidades concretas do mercado de trabalho, facilitando uma integração mais eficaz da força laboral local. As descobertas deste estudo oferecem potenciais ideias aplicáveis a outras regiões urbanas em países em desenvolvimento que enfrentam desafios similares de desemprego. E deste modo, o estudo transcende o mero interesse académico, apresentando potencial para contribuir significativamente para o desenvolvimento económico e social de Moçambique.

1.5 Estrutura do Trabalho

O presente trabalho é composto por cinco (5) capítulos, nomeadamente:

- **Capítulo 1:** Faz-se uma breve introdução do tema, apresenta-se os objectivos, a formulação do problema, assim como as razões que levam a execução da pesquisa.
- **Capítulo 2:** Reserva-se a apresentação de conceitos e fundamentação teórica relacionada a factores que influenciam a ocorrência do desemprego na Cidade e Província de Maputo, e dos modelos estatísticos aplicados.
- **Capítulo 3:** Descreve a metodologia usada, a base de dados usada e a respectiva fonte de dados, as variáveis e a técnica estatística aplicada na análise de dados.

- **Capítulo 4:** Apresentam-se resultados das análises estatísticas feitas e faz-se a interpretação e discussão dos mesmos.

  Capítulo 5: Descreve as principais conclusões do estudo, bem como algumas recomendações.

CAPÍTULO 2

REVISÃO DE LITERATURA

2.1 Conceitos Fundamentais

2.1.1 Definição de Desemprego

A definição de desemprego é um tema central na análise do mercado de trabalho, sendo objecto de constante debate e refinamento na literatura económica e nas práticas estatísticas internacionais. A Organização Internacional do Trabalho (OIT) desempenha um papel fundamental na padronização desta definição, visando facilitar comparações internacionais e orientar políticas públicas (International Labour Organization [ILO], 2013). Segundo a definição padrão da OIT, uma pessoa é considerada desempregada se estiver em idade activa, não estiver trabalhando, estiver disponível para trabalhar e estiver activamente procurando emprego nas últimas quatro semanas (ILO, 2013). Esta definição, amplamente adoptada, serve como base para a maioria das estatísticas oficiais de desemprego em todo o mundo.

No entanto, a abrangência e a adequação desta definição têm sido questionáveis, especialmente em contextos de economias em desenvolvimento e em mercados de trabalho com alto grau de informalidade. Brandolini et al. (2006) argumentam que a definição padrão pode subestimar o verdadeiro nível de desemprego, ao não capturar adequadamente os trabalhadores desalentados ou aqueles em situação de subemprego. Estes autores propõem uma abordagem mais ampla, que considere diferentes graus de ligação ao mercado de trabalho.

A complexidade na mensuração do desemprego é ainda mais evidente quando se consideram as diferentes metodologias empregadas pelos países. Por exemplo, o Canadá e os Estados Unidos, apesar de utilizarem definições semelhantes baseadas nas directrizes da OIT, apresentam diferenças sutis em suas pesquisas de força de trabalho. Uma dessas diferenças é que, o Canadá inclui pessoas a partir de 15 anos nas estatísticas de emprego, enquanto os EUA começam a partir de 16 anos (Statistics Canada, 2015). Estas diferenças metodológicas ressaltam a importância de uma compreensão aprofundada dos métodos de colecta e análise de dados ao comparar taxas de desemprego entre países.

Ademais, a OIT reconhece as limitações da definição padrão e tem trabalhado para desenvolver medidas complementares que capturem de forma mais abrangente as realidades do mercado de trabalho. O conceito de "potencial força de trabalho", por exemplo, foi introduzido para incluir pessoas que desejam trabalhar, mas não atendem a todos os critérios da definição padrão de desemprego (ILO, 2018). Esta abordagem visa capturar nuances importantes, como o subemprego por insuficiência de horas e a potencial força de trabalho, proporcionando uma visão mais completa da subutilização da força de trabalho.

Para o caso de Moçambique a definição do desemprego da OIT, sofre algumas alterações, sendo considerado desempregado quem no período de referência estavam na situação de população desempregada segundo OIT, incluindo pessoas nas condições como: Trabalhadores ocasionais, Trabalhadores por conta própria sem empregados e sem trabalho regular, Trabalhadores familiares sem remuneração e sem trabalho regular, Trabalhadores familiares sem remuneração que não trabalharam no período de referência (INE,2023).

É importante notar que a definição e mensuração do desemprego não são apenas exercícios académicos ou estatísticos, mas têm implicações significativas para a formulação de políticas públicas e a compreensão da saúde económica de um país. Como destacado por Hussmanns (2007), a precisão na definição e medição do desemprego é crucial para o desenvolvimento de políticas eficazes de emprego e para a avaliação do impacto de intervenções económicas.

Enquanto a definição padrão da OIT continua sendo a referência principal para estatísticas internacionais de desemprego, o debate académico e as práticas estatísticas estão em constante evolução. A busca por métodos mais abrangentes e nuançados de mensuração do desemprego reflecte o reconhecimento da complexidade e diversidade dos mercados de trabalho globais, especialmente em economias em desenvolvimento como Moçambique.

2.1.2 Tipos de Desemprego

Os tipos de desemprego são classificados de acordo com suas causas e características, oferecendo uma visão mais detalhada sobre as dinâmicas do mercado de trabalho. Compreender essas categorias é essencial para a formulação de políticas eficazes que visem mitigar o desemprego e seus impactos socioeconómicos.

Um dos tipos mais comuns é o desemprego cíclico, que está directamente relacionado às flutuações económicas. Durante períodos de recessão, a demanda agregada por bens e serviços diminui, levando a uma redução na produção e, consequentemente, a cortes de empregos (Yánez Contreras & Cano Hernández, 2011). Este tipo de desemprego é, portanto, temporário e tende a diminuir quando a economia se recupera.

Outro tipo é o desemprego estrutural, que ocorre quando há um desajuste entre as habilidades dos trabalhadores e as necessidades do mercado de trabalho. Mudanças tecnológicas e a globalização são factores que contribuem para o desemprego estrutural. Este tipo de desemprego é mais persistente e requer intervenções, como programas de requalificação profissional, para ajudar os trabalhadores a se adaptarem às novas demandas do mercado (Kapur, 2022).

O desemprego friccional refere-se a um fenómeno que ocorre quando uma parcela da População Economicamente Activa (PEA) se encontra temporariamente desempregada durante o processo de escolha ou mudança de emprego, em um contexto onde teoricamente existe uma vaga disponível para cada trabalhador desempregado, mas devido a imperfeições no mercado de trabalho, essas vagas não são preenchidas imediatamente (Zurrón Ocio, 1995). Este tipo de desemprego pode ser considerado natural e, até certo ponto, inevitável em uma economia dinâmica.

Além destes, existe o desemprego sazonal caracterizado pela ausência de trabalho durante determinadas épocas do ano, afectando principalmente trabalhadores agrícolas e rurais. Este tipo de desemprego ocorre em períodos específicos onde a produtividade é baixa, enquanto nos períodos de alta produtividade os trabalhadores mantêm-se empregados. Uma característica importante é que os trabalhadores, conscientes desta sazonalidade, costumam fazer poupanças durante os períodos de trabalho para sustentar-se durante os períodos de desemprego (Kapur, 2022).

2.1.3 População Economicamente Activa

A População Economicamente Activa (PEA) representa um indicador fundamental para a análise do mercado de trabalho, abrangendo pessoas em idade activa disponíveis para a

produção de bens e serviços na economia. Em Moçambique, a PEA compreende as pessoas de 15 anos ou mais, que trabalham e as que procuram activamente um emprego (INE, 2019).

De acordo com Ferrão et al. (2018), em Moçambique, uma característica marcante da População Economicamente Activa é a predominância da agricultura de subsistência. Os autores destacam que aproximadamente 80% da população depende directamente da agricultura como principal fonte de subsistência, sendo que 73% vive em áreas rurais, o que influencia directamente os padrões de emprego e a produtividade do trabalho.

De acordo com UN Women (2024), globalmente as mulheres tem uma participação menor que os homens na força de trabalho, isto é, a PEA é maioritariamente constituída por Homens do que por mulheres.

2.1.4 Mercado de Trabalho em Moçambique

O mercado de trabalho em Moçambique apresenta características particulares que reflectem os desafios de uma economia em desenvolvimento. Segundo o Banco Mundial (2018), o país enfrenta um dos crescimentos mais acelerados da força de trabalho entre os países da Associação Internacional de Desenvolvimento (IDA), com aproximadamente 500.000 jovens entrando no mercado de trabalho anualmente. Este crescimento exponencial da força laboral impõe pressões significativas sobre a capacidade de absorção do mercado formal.

Na região subsaariana, como é o caso de Moçambique, observa-se uma crescente necessidade de adaptação da População às novas demandas do mercado de trabalho, especialmente em áreas urbanas, onde a transformação digital está remodelando fundamentalmente as habilidades necessárias para que as pessoas possam acessar mercados, operar fábricas ou gerenciar seus próprios negócios. Este processo tem gerado desafios para a qualificação e inserção profissional da população economicamente activa (IFC & L.E.K. Consulting, 2019).

O sector informal em Moçambique representa uma parte substancial da economia do país, empregando aproximadamente 80% da força de trabalho, principalmente na agricultura e no trabalho autónomo informal, com apenas 6% da força de trabalho coberta pela seguridade social (Aga et al., 2021). Este cenário reflecte os desafios estruturais enfrentados pelo país em promover a formalização do trabalho e garantir condições dignas de emprego para população.

De acordo com o Ministério do Género, Criança e Acção Social (MGCAS, 2022), considerando como sector formal as empresas registadas com cumprimento de obrigações fiscais e laborais, apenas 4% das mulheres conseguem inserção neste mercado. A maioria desenvolve actividades no sector informal, caracterizado pela ausência de registo empresarial e protecção social, incluindo agricultura de subsistência, comércio ambulante e pequenos negócios familiares.

O Mercado de trabalho na Cidade de Maputo enfrenta desafios estruturais complexos que afectam significativamente a empregabilidade juvenil. Entre os obstáculos mais críticos, destacam-se a corrupção no processo de contratação, através de "padrinhos" que vendem empregos, a falta de experiência prática dos jovens e uma formação profissional que nem sempre está alinhada às necessidades actuais do mercado (Fundação Fé e Cooperação [FEC], 2023).

2.2 Factores Associados ao Desemprego

O desemprego é um fenómeno complexo que reflecte a interacção de diversos factores económicos, sociais e demográficos. As taxas de desemprego variam significativamente em função de características individuais, como escolaridade, idade, género, posição no domicílio entre outros factores (Oliveira, Scorzafave, & Pazello, 2009).

Msigwa e Kipesha (2013), ao analisarem os factores associados ao Desemprego, identificaram o género como um factor crucial nas disparidades de acesso ao mercado de trabalho. Os autores evidenciaram uma significativa desigualdade de género, onde mulheres tem menos chance de conseguir um emprego em comparação aos homens.

Cunha et al. (2011) enfatizam a influência da escolaridade como factor determinante, demonstrando que indivíduos com níveis educacionais mais elevados apresentam menores taxas de desemprego. No entanto, os autores também destacam um paradoxo interessante: em algumas regiões metropolitanas, jovens com educação superior enfrentam dificuldades significativas na inserção no mercado de trabalho.

Borchers et al. (2022) analisaram o impacto do estado civil na empregabilidade, revelando que pessoas casadas apresentam menor probabilidade de desemprego em comparação com

solteiros. Os autores atribuem este fenómeno à maior pressão por estabilidade financeira e responsabilidades familiares.

A posição no agregado familiar emerge como factor crucial nos estudos de Silva et al. (1999), que demonstram que chefes de família apresentam menor taxa de desemprego em comparação com aqueles que ocupam outras posições no núcleo familiar.

Yánez Contreras e Cano Hernández (2011) abordam a questão do tamanho do agregado familiar, evidenciando que famílias mais numerosas tendem a apresentar maior vulnerabilidade ao desemprego, especialmente em contextos urbanos onde o custo de vida é mais elevado.

Duarte (2021) demonstra que a idade é um determinante crítico do desemprego, com jovens entre 18 e 24 anos enfrentando taxas de desemprego superiores em comparação a pessoas mais velhas. O autor identifica um padrão em forma de "U" invertido na relação idade-emprego, com as chances de empregabilidade aumentando gradualmente até os 45 anos e depois declinando.

Costa e Cunha (2010) exploram a dimensão geográfica do desemprego, revelando disparidades significativas entre áreas urbanas e rurais. Seus resultados indicam que residentes em áreas metropolitanas enfrentam maior competição por vagas e, consequentemente, maiores taxas de desemprego.

A Companhia de Planejamento do Distrito Federal (2021) traz uma importante contribuição ao analisar o impacto da deficiência nas taxas de desemprego. O estudo revela que pessoas com deficiência enfrentam taxas de desemprego significativamente mais altas, mesmo em contextos onde existem políticas de inclusão.

2.3 Modelos Lineares Generalizados

No ambiente científico, pesquisadores frequentemente se deparam com o desafio de compreender e quantificar as relações entre variáveis de interesse. Esta necessidade perpassa praticamente todas as áreas do conhecimento, desde estudos epidemiológicos até análises econométricas. A busca por estas relações tem levado os investigadores a recorrerem a diferentes técnicas estatísticas, sendo a análise de regressão uma das abordagens mais

utilizadas. O processo de modelagem estatística permite não apenas descrever as relações entre variáveis, mas também realizar previsões e tomar decisões baseadas em evidências. Neste contexto, diferentes estruturas de modelagem foram desenvolvidas ao longo do tempo, buscando atender às especificidades de cada tipo de dado e às diferentes naturezas das relações entre as variáveis em estudo.

A história dos Modelos Lineares Generalizados (MLG) remonta ao início do século XX, quando os estatísticos trabalhavam principalmente com o modelo linear normal. Durante décadas, o modelo linear normal foi o paradigma dominante na análise de regressão, baseando-se nos pressupostos de normalidade, homogeneidade de variância e linearidade (McCullagh & Nelder, 1989). No entanto, à medida que problemas mais complexos surgiam em áreas como biologia, medicina e ciências sociais, as limitações do modelo linear normal tornavam-se cada vez mais evidentes.

O modelo linear normal, embora eficiente para variáveis resposta contínuas e normalmente distribuídas, mostrava-se inadequado para análise de dados binários, contagens ou proporções. Estas limitações levaram ao desenvolvimento de técnicas específicas para cada tipo de dado, como a regressão logística para respostas binárias e a regressão de Poisson para dados de contagem. No entanto, estas abordagens eram tratadas como técnicas distintas, sem um framework unificador que permitisse uma compreensão mais ampla de suas conexões teóricas (Lindsey, 1997).

A grande revolução ocorreu em 1972, quando Nelder e Wedderburn publicaram seu trabalho seminal introduzindo os Modelos Lineares Generalizados. Esta contribuição unificou diversos modelos estatísticos existentes sob uma única estrutura teórica, permitindo a análise de variáveis resposta com diferentes distribuições da família exponencial. Os MLG emergiram como uma extensão do modelo linear clássico, mantendo sua simplicidade conceitual, mas expandindo significativamente seu escopo de aplicação (Paula, 2010).

Os Modelos Lineares Generalizados, introduzidos por Nelder e Wedderburn, representam uma extensão dos modelos lineares clássicos. Estes modelos permitem que a variável resposta siga outras distribuições além da normal e proporcionam maior flexibilidade na análise de dados (Nelder & Wedderburn, 1972).

Família Exponencial

A família exponencial de distribuições desempenha um papel central na teoria dos Modelos Lineares Generalizados, fornecendo uma estrutura matemática unificadora para diversas distribuições de probabilidade (Nelder & Wedderburn, 1972; Cordeiro & Demétrio, 2008). Uma variável aleatória Y pertence à família exponencial se sua função densidade de probabilidade (ou função de probabilidade) pode ser expressa na forma canónica (McCullagh & Nelder, 1989):

$$f(y;\theta,\phi) = exp\left\{\frac{[y\theta - b(\theta)]}{\mathrm{a}(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

onde:

θ é o parâmetro canónico, ϕ é o parâmetro de dispersão e a(.), b(.), c(.) são funções reais conhecidas.

Para distribuições pertencentes à família exponencial, algumas propriedades fundamentais são derivadas directamente de sua estrutura:

- Média: $E(Y)=\mu=b'(\theta)$
- Variância: $Var(Y)=a(\phi)b''(\theta)$

onde b'(θ) e b''(θ) são respectivamente, a primeira e segunda derivadas de b(θ).

Os Modelos Lineares Generalizados constituem uma extensão dos modelos lineares clássicos, ampliando significativamente sua aplicabilidade em diferentes contextos de análise estatística (Nelder & Wedderburn, 1972). No modelo linear clássico, a estrutura básica é representada por $Y=X\beta+\epsilon$, onde Y é o vector de respostas, X é a matriz do modelo, β é o vector de parâmetros e ε é o vector de erros aleatórios com distribuição $\epsilon \sim N(0, \sigma^2 I)$ (McCullagh & Nelder, 1989).

A extensão dos modelos lineares clássicos para os MLG ocorre em duas direcções principais. A primeira refere-se à distribuição da variável resposta, que deixa de ser restrita à distribuição normal e passa a abranger qualquer distribuição pertencente à família exponencial, incluindo as distribuições normal, binomial, Poisson e gama, entre outras. A segunda direcção diz respeito à função de ligação, que no modelo clássico estabelecia uma relação directa entre

$E(Y)=\mu$ e o preditor linear $\eta=X\beta$. Nos MLG, introduz-se uma função de ligação g(·) que relaciona $\eta=g(\mu)$ onde g(·) é uma função monótona e diferenciável (Paula, 2010).

Os MLG são caracterizados por três componentes fundamentais que estabelecem sua estrutura matemática (Cordeiro & Demétrio, 2008). O componente aleatório corresponde à variável resposta Y, que segue uma distribuição da família exponencial, com observações independentes $Y_1,..., Y_n$ e suas respectivas médias $\mu_1,\ldots\mu_n$. A função densidade de probabilidade pode ser expressa como:

$$f(y;\theta,\phi) = exp\{\phi^{-1}\ [y\theta - b(\theta)]\ + c(y,\phi)\} \quad (2.2)$$

O componente sistemático, também conhecido como preditor linear, é definido por $\eta=X\beta$, onde η é o preditor linear, X é a matriz de delineamento (n × p) contendo os valores das variáveis explicativas, e β é o vector (p × 1) de parâmetros desconhecidos. Este componente pode ser expresso de forma detalhada como:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \ \ , i= 1,\ldots,n \quad (2.3)$$

A função de ligação g (·), terceiro componente, estabelece a relação entre o valor esperado do componente aleatório e o preditor linear através da equação $g(\mu)=\eta=X\beta$. Esta função deve ser monótona e diferenciável.

Algumas funções de ligação comumente utilizadas incluem:

- Ligação identidade: $g(\mu)= \mu$
- Ligação logarítmica: $g(\mu)= \log \mu$
- Ligação logística: $g(\mu)= \log \frac{\mu}{1-\mu}$
- Ligação recíproca: $g(\mu)= \frac{1}{\mu}$

2.3.1 Estimação dos Parâmetros

A estimação dos parâmetros nos Modelos Lineares Generalizados é tradicionalmente realizada através do método da máxima verossimilhança (McCullagh & Nelder, 1989). Este método fornece estimadores com propriedades assimptóticas desejáveis, como consistência e eficiência.

O logaritmo da função de verossimilhança para uma única observação, considerando a forma canónica da família exponencial, é dado por:

$$l(\theta_i, \phi;\ y_i) = \ \phi^{-1}\ [y_{i\theta_i} - \ b(\theta_i)] + \ c(y_i, \phi) \tag{2.4}$$

Para uma amostra de n observações independentes, o logaritmo da função de verossimilhança total é obtido pela soma das contribuições individuais (Cordeiro & Demétrio, 2008):

$$l(\theta, \phi;\ y) = \ \phi^{-1}\ \sum_{i=1}^{n} [y_{i\theta_i} - \ b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi) \tag{2.5}$$

As equações de verossimilhança são obtidas derivando L(θ,φ;y) em relação a cada componente do vector de parâmetros β e igualando a zero. Como destacado por Paula (2010), este sistema de equações geralmente não possui solução analítica, necessitando de métodos iterativos para sua resolução. O método mais utilizado é o dos Escores de Fisher, também conhecido como método de Newton-Raphson modificado.

O processo iterativo pode ser expresso através da equação:

$$\beta^{(m+1)} \ = \ \beta^{(m)} \ + \ (X^T W^{(m)} X)^{-1} X^T W^{(m)} \mathrm{z}^{(m)} \tag{2.6}$$

onde:

$\beta^{(m+1)}$ é a estimativa actualizada do vector de parâmetros;

$\beta^{(m)}$ é a estimativa actual da iteração m;

$W^{(m)}$ é a matriz de pesos da iteração m;

$\mathrm{z}^{(m)}$ é a variável dependente modificada da iteração m.

A matriz de pesos W é diagonal com elementos dados por:

$$W_i = \frac{1}{Var(Y_i)[\frac{d\eta_i}{d\mu_i}]^2} \tag{2.7}$$

onde $\frac{d\eta_i}{d\mu_i} = [g'_{(\mu_i)}]^{-1}$, sendo $g'_{(\mu_i)}$ a derivada da função de ligação.

Como observado por Dobson e Barnett (2018), o processo iterativo continua até que seja atingido um critério de convergência pré-estabelecido, geralmente baseado na diferença relativa entre estimativas sucessivas.

A matriz de informação de Fisher observada é dada por:

$$\mathrm{I}(\beta) = X^T W X \tag{2.8}$$

e pode ser utilizada para obter os erros padrão aproximados das estimativas dos parâmetros através de:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{[I(\hat{\beta})^{-1}]_{jj}} \tag{2.9}$$

2.3.2 Testes de Hipóteses em MLG

No contexto dos Modelos Lineares Generalizados, os testes de hipóteses são fundamentais para avaliar a significância dos parâmetros do modelo e a adequação do ajuste. Segundo Cordeiro e Demétrio (2008), existem três testes principais amplamente utilizados: o teste de Wald, o teste da razão de verossimilhança (TRV) e o teste escore.

Para testar hipóteses sobre os parâmetros β do modelo, consideramos geralmente a hipótese nula $H_0: C\beta = \varepsilon_0$ contra a hipótese alternativa $H_1: C\beta \neq \varepsilon_0$, onde C é uma matriz q × p de posto q ≤ p e ε_0 é um vetor conhecido de dimensão q (Paula, 2010).

Teste de Wald

O teste de Wald é baseado na distribuição normal assimptótica dos estimadores de máxima verossimilhança. A forma geral da estatística de Wald é dada por:

$$W = (C\hat{\beta} - \varepsilon_0)^T [C(X^T \widehat{W} X)^{-1} C^T]^{-1} (C\hat{\beta} - \varepsilon_0) \tag{2.10}$$

Para o caso particular de testar a significância de um único parâmetro H_0: $\beta_j = 0$, a estatística de Wald reduz-se a:

$$W_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \tag{2.11}$$

Teste da Razão de Verossimilhança (TRV)

O TRV compara os logaritmos das funções de verossimilhança maximizadas sob as hipóteses nula e alternativa. A estatística do teste é:

$$\lambda = 2[l(\hat{\beta}) - l(\tilde{\beta})] \tag{2.12}$$

onde $l(\hat{\beta})$ é o logaritmo da verossimilhança maximizada sob o modelo completo e $l(\tilde{\beta})$ é o logaritmo da verossimilhança maximizada sob H_0. Sob condições de regularidade e sob H_0, λ tem distribuição assimptótica qui-quadrado com q graus de liberdade (Dobson & Barnett, 2018).

Teste Escore

O teste escore, também conhecido como teste de multiplicador de Lagrange, é baseado no vector escore avaliado sob H_0. A estatística do teste é:

$$S_R = U(\tilde{\beta})^T [I(\tilde{\beta})]^{-1} U(\tilde{\beta}) \tag{2.13}$$

onde $U(\tilde{\beta})$ é o vector escore e $I(\tilde{\beta})$ é a matriz de informação de Fisher. Sob H_0, S_R tem distribuição assimptótica qui-quadrado com q graus de liberdade (Turkman & Silva, 2000).

2.4 Regressão Logística

A regressão logística é uma técnica estatística amplamente utilizada em diversas áreas do conhecimento, como saúde, ciências sociais, economia e engenharia, devido à sua capacidade de modelar relações entre uma variável dependente categórica e uma ou mais variáveis independentes.

Esta metodologia foi inicialmente desenvolvida como uma extensão da regressão linear para lidar com variáveis dependentes categóricas. Sua popularidade deve-se à flexibilidade, simplicidade e robustez que oferece na análise de dados categóricos, especialmente em situações onde os pressupostos clássicos da regressão linear, como normalidade e homogeneidade de variância, não são atendidos (Peng et al., 2002).

O principal objectivo da regressão logística é modelar a probabilidade de ocorrência de um evento de interesse em função de variáveis explicativas. Para isso, utiliza-se a função logit, que transforma a probabilidade de ocorrência em uma escala linear, permitindo que o modelo seja ajustado por métodos de máxima verossimilhança. Essa abordagem é mais adequada para variáveis dependentes categóricas do que os métodos tradicionais de mínimos quadrados utilizados na regressão linear (Agresti, 2013).

Uma das vantagens da regressão logística é que as previsões geradas estão sempre dentro do intervalo válido de probabilidade [0,1], o que garante interpretações consistentes e realistas dos resultados (Kleinbaum & Klein, 2010).

Diversos estudos destacam a aplicabilidade da regressão logística em contextos onde a variável dependente é binária, como a presença ou ausência de uma doença, a ocorrência ou não de um evento, ou a classificação de indivíduos em dois grupos distintos.

Peng et al. (2002) enfatizam que a regressão logística é particularmente útil para análises de predição e classificação, permitindo que pesquisadores identifiquem os factores que influenciam a probabilidade de ocorrência de um evento. Além disso, a técnica não exige que as variáveis independentes sejam normalmente distribuídas ou que apresentem variâncias homogéneas, o que amplia significativamente seu campo de aplicação (Hosmer et al., 2013).

A interpretação dos coeficientes estimados pelo modelo é outro aspecto amplamente discutido na literatura. Os coeficientes da regressão logística são geralmente interpretados em termos de

razões de chances (*odds ratio*), que representam o efeito multiplicativo de uma unidade de aumento na variável independente sobre a probabilidade relativa do evento de interesse ocorrer. Essa interpretação é especialmente útil em áreas como epidemiologia, onde as razões de chances são frequentemente utilizadas para quantificar associações entre factores de risco e desfechos clínicos. Além disso, a utilização de razões de chances facilita a comunicação dos resultados para públicos não técnicos, tornando a técnica acessível para formuladores de políticas e tomadores de decisão (Sperandei, 2014).

Outro ponto relevante é a capacidade da regressão logística de lidar com variáveis independentes categóricas. Essas variáveis são geralmente codificadas como variáveis dummy, permitindo que o modelo capture as diferenças entre os grupos categóricos de forma eficiente. Essa abordagem é particularmente útil em estudos que envolvem variáveis qualitativas, como género, escolaridade ou estado civil (Agresti, 2013). Além disso, a técnica permite explorar interacções entre variáveis independentes, possibilitando a identificação de efeitos moderadores e mediadores que podem influenciar a relação entre as variáveis explicativas e a variável dependente (Hair Jr. et al., 2009).

A literatura também destaca as extensões da regressão logística para lidar com variáveis dependentes politômicas e ordinais. A regressão logística multinomial, por exemplo, é utilizada quando a variável dependente possui mais de duas categorias não ordenadas, enquanto a regressão logística ordinal é aplicada em situações onde as categorias da variável dependente possuem uma ordem natural. Essas extensões ampliam ainda mais as possibilidades de aplicação da técnica, permitindo que ela seja utilizada em uma ampla gama de contextos de pesquisa (Hosmer et al., 2013).

Apesar de suas vantagens, a regressão logística também apresenta algumas limitações que são frequentemente discutidas na literatura. Uma delas é a suposição de independência das observações, que pode ser violada em estudos com dados hierárquicos ou longitudinais. Para lidar com essa limitação, técnicas mais avançadas, como modelos de regressão logística hierárquica ou modelos de efeitos mistos, têm sido desenvolvidas e aplicadas em contextos onde as observações estão agrupadas ou correlacionadas (Raudenbush & Bryk, 2002).

Em síntese, a regressão logística é uma ferramenta estatística versátil e amplamente aplicada na análise de dados categóricos. Sua capacidade de modelar probabilidades, lidar com variáveis de diferentes naturezas e fornecer interpretações claras dos resultados torna-a uma escolha

popular entre pesquisadores de diversas disciplinas. No entanto, como qualquer técnica, ela possui limitações que devem ser consideradas e abordadas de forma adequada para garantir a validade e a confiabilidade dos resultados. A literatura destaca tanto os pontos fortes quanto as limitações da regressão logística, fornecendo um panorama abrangente de suas aplicações e possibilidades de aprimoramento.

Pressupostos da Regressão Logística

A regressão logística, embora seja uma técnica estatística flexível e amplamente utilizada, possui pressupostos que devem ser atendidos para garantir a validade e a confiabilidade dos resultados. Diferentemente da regressão linear, a regressão logística não exige que a variável dependente seja contínua ou normalmente distribuída, nem que as variáveis independentes apresentem homogeneidade de variância. No entanto, ela requer que as observações sejam independentes entre si, ou seja, a ocorrência de um evento em uma unidade de análise não deve influenciar a ocorrência do mesmo evento em outra unidade. Essa suposição é fundamental para evitar vieses nas estimativas dos coeficientes e na significância estatística, especialmente em estudos com dados agrupados ou repetidos (Hair Jr. et al., 2009).

A regressão logística pode ser sensível à presença de multicolinearidade entre as variáveis independentes, o que pode distorcer as estimativas dos coeficientes e dificultar a interpretação dos resultados, por isso esta técnica tem como um dos pressupostos a ausência de multicolinearidade perfeita entre as variáveis independentes (Meyers et al., 2006).

Adicionalmente, a regressão logística pressupõe que a variável dependente seja categórica, geralmente binária, representando dois estados mutuamente exclusivos, como "sucesso" e "fracasso". Para variáveis dependentes com mais de duas categorias, é necessário utilizar extensões da regressão logística, como a regressão logística multinomial ou ordinal. Também é importante que as categorias da variável dependente sejam claramente definidas e mutuamente excludentes, para evitar ambiguidades na classificação dos casos (Agresti, 2013).

Por último, a adequação do modelo deve ser avaliada por meio de testes de qualidade de ajuste, como o teste de Hosmer-Lemeshow, que avalia se o modelo obtido pode explicar adequadamente os dados observados, dividindo os dados em g grupos segundo as probabilidades estimadas. A verificação da qualidade de ajuste é uma etapa fundamental para

garantir que o modelo seja adequado aos dados e que as conclusões sejam confiáveis (Hosmer et al., 2013).

Comparada à análise discriminante, a regressão logística apresenta algumas vantagens importantes, especialmente no que diz respeito aos pressupostos. Enquanto a análise discriminante depende da suposição de normalidade multivariada das variáveis independentes dentro de cada grupo da variável dependente, a regressão logística não exige essa condição, tornando-a mais robusta em situações onde a normalidade não é atendida. Além disso, a análise discriminante pressupõe homogeneidade das matrizes de covariância entre os grupos, um requisito que, quando violado, pode comprometer a precisão das classificações. Por outro lado, a regressão logística é menos sensível a essas violações e pode ser aplicada com maior flexibilidade, especialmente em contextos com variáveis independentes de diferentes escalas ou distribuições (Hair Jr. et al., 2009).

Modelo de Regressão Logística

O principal objectivo do modelo de regressão logística é analisar como um conjunto de variáveis explicativas influencia a probabilidade de ocorrência do evento. Diferentemente da regressão linear, que prevê valores contínuos, a regressão logística binária transforma o espaço linear das variáveis explicativas em um espaço probabilístico, garantindo que os valores estimados estejam no intervalo entre 0 e 1 (Agresti, 2013)

O modelo de regressão logística binária começa com uma estrutura semelhante à regressão linear. Segundo Fávero e Belfiore (2017), define-se um vector de variáveis explicativas, com respectivos parâmetros estimados, da seguinte forma:

$$Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_k X_{ki} \qquad (2.14)$$

Onde:

Z_i é a combinação linear das variáveis explicativas, conhecida como logito;

β_0 é o intercepto do modelo;

β_k são os parâmetros estimados de cada variável explicativa;

X_k são as variáveis explicativas (métricas ou dummies);

i representa cada observação da amostra (i = 1, 2, ..., n, em que n é o tamanho da amostra).

Para compreender a probabilidade de ocorrência do evento, é necessário introduzir o conceito de razão de chances (odds), que é definido como a razão entre a probabilidade de ocorrência do evento p_i e a probabilidade de não ocorrência (1 - p_i):

$$chances(odds)_{Y_{i=1}} = \frac{p_i}{1 - p_i} \tag{2.15}$$

A regressão logística binária define o logito Z como o logaritmo natural da chance, de modo que:

$$\ln (chances_{Y_{i=1}}) = Z_i \tag{2.16}$$

de onde vem que:

$$\ln (\frac{p_i}{1 - p_i}) = Z_i \tag{2.17}$$

Como o intuito é definir uma expressão para a probabilidade de ocorrência do evento em estudo, em função do logito, podemos matematicamente isolar P; a partir da expressão (2.16), da seguinte maneira:

$$\frac{p_i}{1 - p_i} = e^{Z_i} \tag{2.18}$$

$$p_i = (1 - p_i)\, e^{Z_i} \tag{2.19}$$

$$p_i(1+e^{Z_i}) = e^{Z_i} \tag{2.20}$$

E, portanto, temos que:

Probabilidade de ocorrência do evento:

$$p_i=\frac{e^{Z_i}}{(1+e^{Z_i})}=\frac{1}{(1+e^{-Z_i})} \tag{2.21}$$

A partir das expressões (2.14) e (2.21), podemos definir a expressão geral da probabilidade estimada de ocorrência de um evento que se apresenta na forma dicotómica para uma observação i da seguinte forma:

$$p_i=\frac{1}{1+e^{-(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots\beta_kX_{ki})}} \tag{2.22}$$

2.4.1 Regressão logística Simples

De acordo com Hosmer e Lemeshow (2013) a regressão logística simples é utilizada quando há apenas uma variável independente X para prever a probabilidade de ocorrência de um evento Y. Nesse caso, a probabilidade de Y=1 é modelada pela seguinte equação:

$$p_{(Y=1|X)}=\frac{1}{1+e^{-(\beta_0+\beta_1X_1)}} \tag{2.23}$$

Sendo β_0e β_1 parametros desconhecidos do modelo.

2.4.2 Regressão logística Múltipla

Quando há mais de uma variável independente, utiliza-se a regressão logística múltipla. Nesse caso, a probabilidade de Y=1 é modelada como uma função logística de uma combinação linear de múltiplas variáveis independentes $X_1, X_2,\ldots\ldots X_{\mathrm{k}}$:

$$p_{(Y=1|X)}=\frac{1}{1+e^{-(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots\beta_kX_{ki})}} \tag{2.24}$$

2.4.3 Estimação dos Parâmetros no Modelo de Regressão Logística Binária

De acordo com Figueira (2006), para ajustar um modelo de regressão logística, é necessário estimar os parâmetros β_0 e β_1 do modelo. No entanto, para isso, utiliza-se o método de máxima verossimilhança.

Segundo Meyer (1980), o método de máxima verossimilhança de β, baseado em uma amostra aleatória $x_1, x_2, \dots, x_n$ é aquele valor de β que torna máxima $L(x_1, x_2, \dots, x_n; \beta)$, considerada como uma função de β para uma dada amostra $x_1, x_2, \dots, x_n$.

A função verossimilhança é definida por:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{Y_i}[(1 - \pi(x_i))^{1-Y_i}] \, , \beta \in R^2 \tag{2.25}$$

O princípio da máxima verossimilhança é estimar o valor de β que maximiza $L(\beta)$.

Aplicando o logaritmo natural em ambos lados da equação, obtemos a função log-verosimilhança:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} [Y_i \ln \pi(x_i) + (1 - Y_i)\ln\,(1 - \pi_i)] \tag{2.26}$$

Para encontrar o valor de β que maximiza $l(\beta)$ deriva-se $l(\beta)$ em relação a cada parâmetro (β_0, β_1), obtendo-se duas equações:

$$\frac{\partial l(\beta)}{\partial(\beta_0)} = \sum_{i=1}^{n} [y_i - \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} * e^{\beta_0 + \beta_1 x_i}] \tag{2.27}$$

$$\frac{\partial l(\beta)}{\partial(\beta_1)} = \sum_{i=1}^{n} [y_i x_i - \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} * x_i e^{\beta_0 + \beta_1 x_i}] \tag{2.28}$$

Que igualando a zero, geram o seguinte sistema de equações:

$$\begin{cases} \sum_{i=1}^{n}(y_i - \pi_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - \pi_i) = 0 \end{cases} \tag{2.29}$$

em que $i = 1, 2, \dots, n$ e $\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$

Conforme Belfiore (2015), a regressão logística múltipla é uma extensão de regressão binária. No modelo anterior, temos uma única variável independente. Para o caso onde temos um conjunto p variáveis independentes expressas pelo vector $x^T \equiv (x_1, x_2, \dots, x_p)$. Para o Hosmer e Lemeshow (2013), no modelo de probabilidade múltipla, a probabilidade de sucesso é dada por:

$$\pi_i = P(Y_i = 1|X_i = x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}} = \frac{e^{({x_i}^T\beta)}}{1 + e^{({x_i}^T\beta)}} \tag{2.30}$$

Para estimar os parâmetros da regressão logística múltipla por máxima verosimilhança encontra-se o valor de β que maximiza $l(\beta)$, o que exige um processo iterativo e que faz necessário derivar $l(\beta)$ em relação a cada parâmetro:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} [y_i x_{ij} - \frac{e^{({x_i}^T\beta)}}{1 + e^{({x_i}^T\beta)}} x_{ij}] \tag{2.31}$$

Onde $l(\beta)$ é o logaritmo da função (2.25).

A matriz de covariância dos coeficientes estimados é obtida a partir das derivadas parciais de segunda ordem do logaritmo da função de verossimilhança:

$$\frac{\partial^2 l(\beta)}{\partial {\beta_j}^2} = -\sum_{i=1}^{n} [{x_{ij}}^2 \pi_i(1 - \pi_i)] \tag{2.32}$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} [x_{ij} x_{ik} \pi_i (1 - \pi_i)] \tag{2.33}$$

Onde: $j, k = 0,1,2, \dots, p$ e π_i representa $\pi(x_i)$.

Se for formada uma matriz quadrada de dimensão $(p+1) \times (p+1)$, constituída pelo simétrico dos valores médios dos termos referidos nas equações (2.32) e (2.33), obtém-se $I(\beta)$, a chamada matriz de informação.

A matriz de informação de Fisher é dada por:

$$I(\beta) = -E\left(\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k}\right) = X^T Q X \tag{2.34}$$

Em que, $Q = diag[\pi_i(1 - \pi_i)], i = 1, \dots, n$ e X a matriz de dados. A inversa desta matriz, $[I(\beta)^{-1}]$, é a matriz de variância e covariância das estimativas de máxima verossimilhança dos parâmetros.

CAPÍTULO 3

Material e Métodos

3.1 Material

Para a realização do presente trabalho, utilizou-se uma base de dados secundária obtida no Instituto Nacional de Estatística, referente aos dados do Inquérito sobre Orçamento Familiar (IOF) realizado em 2022.

O Inquérito sobre Orçamento Familiar (IOF) de 2022 foi desenvolvido com base nos dados do IV Recenseamento Geral da População e Habitação de 2017, conduzido pelo Instituto Nacional de Estatística. Este inquérito assegura representatividade a nível nacional, regional, urbano e rural. A amostra foi desenhada em três etapas principais: Na primeira etapa, a amostra foi estratificada por província e por área urbana/rural. Foram seleccionadas 1.496 Unidades Primárias de Amostragem (UPAs) de forma sistemática, com probabilidades iguais. Na segunda etapa, em cada UPA seleccionada, uma única Área de Enumeração (AE) foi escolhida utilizando o método de Probabilidade Proporcional ao Tamanho (PPT), considerando o número de agregados familiares em cada AE. Na terceira etapa, dentro de cada AE, todos os agregados familiares foram listados, e uma amostra foi seleccionada de forma sistemática com probabilidades iguais: 12 agregados familiares em áreas urbanas e 9 em áreas rurais, totalizando uma amostra de 16.992 agregados familiares em todo o país. Portanto, para o estudo em particular, foram seleccionadas apenas duas províncias, nomeadamente: Maputo Província e Maputo Cidade. Para além disso, considerou-se apenas indivíduos economicamente activos (que inclui todas as pessoas com 15 anos de idade ou mais) correspondendo 9761 indivíduos.

O processamento dos dados foi realizado com recurso ao software R versão 4.2.4. Todas as hipóteses foram testadas considerando um nível de significância de 5%, sendo a regra de decisão baseada no p-valor associado à estatística do teste. Este trabalho foi elaborado com recurso ao Microsoft Office Word 365. A Tabela 3.1 apresenta as variáveis usadas neste estudo.

Tabela 3.1: Descrição das variáveis usadas no estudo

Variáveis	Descrição
Idade	Idade do respondente em anos
Sexo	Sexo do respondente (Homem ou Mulher)
Casado	Estado civil do respondente (casado ou não)
Província	Província de residência do respondente
Região	Região onde o respondente reside
Escolaridade	Nível de escolaridade do respondente
Parentesco	Relação de parentesco do respondente
Deficiência	Indica se o respondente tem alguma deficiência
Tamanho do agregado	Indica o número de membros do agregado familiar
Desemprego	Indica se o respondente está desempregado
Weight	Ponderador da amostra para análise estatística

3.2 Métodos

3.2.1 Associação entre as Variáveis do Estudo

De acordo com Agresti e Finlay (2009), aplica-se o teste de independência quando os dados da pesquisa estão de forma de frequências em categorias. Para aplicação deste teste de hipótese, recorre-se à estatística de Qui-quadrado (χ^2) para determinar se duas variáveis são independentes ou se existe uma associação entre elas, utilizando a seguinte estatística:

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2(L-1)(C-1) \tag{3.1}$$

Onde:

O_{ij} − representa o número de casos observados na linha i da coluna j; e

E_{ij} −representa o número de casos esperados na linha i da coluna j.

As hipóteses a serem testadas são:

H_0: As variáveis são independentes, isto é, não há associação entre as variáveis

H_1: As variáveis não são independentes, isto é, há associação entre as variáveis

3.2.2 Métodos de Selecção das Variáveis

Para desenvolver um modelo eficiente com o menor número de covariáveis possível, é crucial definir um plano para seleccionar as covariáveis iniciais que serão testadas, além de adoptar um método que auxilie na escolha e ajuste dessas variáveis (James, Witten, Hastie, & Tibshirani, 2013). Segundo Draper e Smith (1998), a selecção das variáveis do modelo final é baseada em algoritmos que avaliam a importância de cada variável e decidem sua inclusão ou exclusão no modelo. Métodos comuns incluem forward, backward e stepwise.

Neste estudo, foi escolhido o método backward. Ele começa com todas as variáveis candidatas incluídas no modelo e procede com a avaliação da significância estatística de cada uma. A cada passo, a variável com menor contribuição para o modelo (maior p-value) é removida, desde que não seja estatisticamente significativa. Após cada remoção, o modelo é reavaliado para verificar se as variáveis restantes mantêm sua significância estatística. O processo continua até que todas as variáveis presentes no modelo sejam estatisticamente significativas e nenhuma variável adicional precise ser removida.

3.2.3 Critérios para a Selecção do Modelo

A escolha do melhor modelo é um desafio, pois é necessário equilibrar a qualidade do ajuste com a complexidade, geralmente medida pelo número de parâmetros. Mais parâmetros tornam o modelo mais complexo e difícil de interpretar, tornando essencial a selecção do melhor modelo (Burnham & Anderson, 2002).

Critério de Informação de Akaike

O Critério de Informação de Akaike (AIC) foi introduzido por Akaike (1974) para melhorar a selecção de modelos utilizando a divergência de Kullback-Leibler. Ele relaciona a máxima

verossimilhança com essa divergência, criando um critério para estimar a informação perdida quando um modelo específico é utilizado.

O AIC realiza um processo de minimização que não envolve testes estatísticos, sendo expresso em função do desvio do modelo e baseado na função de verossimilhança. Ele oferece uma medida relativa das informações perdidas ao usar um modelo para descrever a realidade. O AIC é definido como:

$$\mathrm{AIC} = -2(\mathrm{LL}) + 2(\mathrm{K}) \tag{3.2}$$

onde K é o número de parâmetros no modelo estatístico e LL é o valor maximizado da função de verossimilhança para o modelo estimado.

O AIC não testa hipóteses, mas é uma ferramenta de selecção de modelos; não envolve significância ou valor-p. Modelos com menor AIC são preferidos.

Critério de Informação Bayesiano

O Critério de Informação Bayesiano (BIC), também conhecido como Critério de Schwarz, foi proposto por Schwarz (1978) como um método de avaliação de modelos baseado na probabilidade a posteriori. O BIC é definido como:

$$\mathrm{BIC} = -2(\mathrm{LL}) + 2(\mathrm{K} * \ln(n)) \tag{3.3}$$

onde n é o número de observações. Modelos com valores baixos de AIC e BIC são considerados melhores.

3.2.4 Teste de Significância dos Parâmetros do Modelo de Regressão Logística Binária

De acordo com Gonzalez (2018), o teste de razão de máxima verossimilhança é utilizado para testar a significância do modelo estimado. Este teste avalia simultaneamente se os coeficientes β são todos nulos, com exceção de β_0. A comparação entre os valores observados e esperados utilizando a função de máxima verossimilhança é expressa pelas seguintes fórmulas:

$$D = -2ln\left[\frac{verossimilhança\ do\ modelo\ ajustado}{verossimilhança\ do\ modelo\ saturado}\right] \tag{3.4}$$

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{\pi_i}}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{\pi_i}}{1-y_i}\right)\right] \tag{3.5}$$

É considerado modelo saturado se contém todas variáveis, enquanto modelo ajustado é considerado ao modelo apenas com variáveis desejadas ao estudo. A função D, também pode ser chamada de *deviance* (desvio), sempre é positivo e quanto menor, melhor é o ajuste do modelo.

No entanto, testa-se as seguintes hipóteses:

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0\ (O\ modelo\ não\ é\ adequado)$

$H_1: Existe\ pelo\ menos\ um\ \beta_j \neq 0\ i = 1,2,3,\ldots,p\ (O\ modelo\ é\ adequado)$

Para avaliar a significância de uma variável independente, compara-se o valor de D nas situações em que a variável está presente e ausente na equação. A variação esperada no valor de D devido à inclusão da variável independente no modelo é determinada por:

$$G = D(\frac{modelo\ sem\ variável}{Modelo\ com\ variável}) \tag{3.6}$$

Teste de Wald

De acordo com Figueira (2006), geralmente o teste de Wald é também utilizado na regressão logística para a determinação da significância estatística dos coeficientes do modelo estimado. Ele fornece a significância estatística para cada coeficiente estimado, de forma que os testes de hipóteses podem ser realizados exactamente como são feitos na regressão múltipla.

O teste de Wald é obtido comparando a estimativa de máxima verossimilhança de um coeficiente com a estimativa do seu erro padrão:

$$W_j = \frac{\beta_j}{EP(\beta_j)} \tag{3.7}$$

Onde $EP(\beta_j)$ é o erro padrão de β_j. O teste de hipótese referente a cada parâmetro é:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

3.2.5 Qualidade de Ajuste na Regressão Logística Binária

Hosmer e Lemeshow (2013), propuseram um teste de ajustamento muito utilizado na regressão logística, com o objectivo de testar a qualidade do ajuste do modelo. O teste de Hosmer e Lemeshow avalia se o modelo obtido pode explicar adequadamente os dados observados, dividindo os dados em *g* grupos segundo as probabilidades estimadas.

O teste considera as seguintes hipóteses:

H_0: Não existem diferenças significativas entre os resultados previstos e observados,
H_1: Existem diferenças significativas entre os resultados previstos e observados.

A estatística do teste ($\hat{C}$) é dada por:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_{1k} - E_{1k})^2}{(E_{1k})} + \frac{(O_{0k} - E_{0k})^2}{(E_{0k})} \tag{3.8}$$

Onde:

O_{1k} número observado de eventos (sucessos, Y=1) no grupo k;

O_{0k} número observado de não-eventos (fracassos, Y=0) no grupo k;

E_{1k} número esperado de eventos (sucessos, Y=1) no grupo k;

E_{0k} número esperado de não-eventos (fracassos, Y=0) no grupo k;

g número de grupos;

k índice do grupo (k = 1, 2, ..., g).

Multicolinearidade

A multicolinearidade é uma condição em que duas ou mais variáveis independentes em um modelo de regressão estão altamente correlacionadas, o que pode afectar a precisão das estimativas dos coeficientes. Na regressão logística binária, assim como em outros tipos de regressão, a presença de multicolinearidade pode inflar as variâncias dos coeficientes, tornando difícil determinar a contribuição individual de cada variável para o modelo (Hosmer & Lemeshow, 2013).

Midi, Sarkar e Rana (2010) recomendam o uso do Factor de Inflação da Variância (VIF) e da Tolerância para diagnosticar multicolinearidade. Uma tolerância abaixo de 0.1 indica um problema de multicolinearidade significativa, enquanto valores de VIF acima de 10 são tradicionalmente problemáticos, embora em regressão logística valores acima de 2.5 já mereçam atenção. Nesses casos, recomenda-se a remoção ou combinação de variáveis correlacionadas.

3.2.6 Interpretação dos Parâmetros do Modelo

Após o ajuste do modelo de regressão logística binária aos dados, o próximo passo passa por realizar uma interpretação dos coeficientes estimados.

De acordo com Hosmer & Lemeshow (2013), os coeficientes estimados das variáveis independentes indicam a taxa de variação de uma função da variável dependente para cada unidade de mudança na variável independente. Essa interpretação envolve dois aspectos principais: a determinação da relação funcional entre a variável dependente e a variável independente e a definição adequada da unidade de mudança da variável independente.

Para o primeiro aspecto, é necessário identificar qual função da variável dependente resulta em uma relação linear com a variável independente, ou seja, determinar a função de ligação (*link function*). No caso da regressão logística, essa função de ligação é a transformação logit, conforme representada na seguinte equação:

$$g(x) = ln\left\{\frac{\pi(x)}{[1-\pi(x)]}\right\} = \beta_0 + \beta_1 x \tag{3.9}$$

Nessa abordagem, os coeficientes estimados indicam a variação no logit resultante de um aumento unitário na variável independente. Para que a interpretação de um coeficiente no modelo de regressão logística seja significativa, é essencial que a diferença entre dois logits tenha um significado claro. Um dos parâmetros mais amplamente utilizados para essa interpretação é o *odds ratio* (OR). No entanto, a forma como os coeficientes são interpretados pode variar dependendo da natureza da variável independente considerada no modelo.

Variável Dicotómica

Qualquer variável independente com dois valores distintos, diz-se $x = a$ versus $x = b$, é a diferença entre os *logits* estimados nesses dois valores. Nesse caso, o *odds ratio* é estimado utilizando a seguinte expressão:

$$\widehat{OR}(a,b) = e^{\widehat{\beta_1}(a-b)} \tag{3.10}$$

O odds ratio representa a medida de associação mais amplamente utilizada em regressão logística. Para uma variável dicotómica codificada como 0 e 1, a relação entre o odds ratio e o coeficiente de regressão é:

$$OR = e^{\beta_1} \tag{3.11}$$

Variável Policotómica

Para o caso de a variável independente ter mais de dois valores distintos, essa variável deve ser recodificada como variáveis dummy, assumindo os valores 0 ou 1 para que sua interpretação seja viável. Se a variável nominal em análise possuir *m* categorias, será necessário criar *m−1* variáveis *dummy* correspondentes a essas categorias, sendo uma categoria escolhida como referência. O cálculo do *odds ratio* segue o mesmo procedimento aplicado a variáveis dicotómicas.

Variável Contínua

Quando um modelo de regressão logística inclui uma variável independente contínua, a interpretação dos coeficientes estimados depende da forma como essa variável foi incorporada

ao modelo e da unidade de medida utilizada. Supondo que o *logit* tenha uma relação linear com a variável x, a equação do *logit* pode ser expressa como $g(x) = \beta_0 + \beta_1 x$. Nesse contexto, o coeficiente β_1 representa a variação no logaritmo natural da razão de chances quando x aumenta em uma unidade, ou seja, $\beta_1 = g(x+1) - g(x)$, independentemente do valor específico de x.

Para tornar essa interpretação mais informativa no caso de uma variável contínua, pode-se estabelecer um critério para avaliar a mudança no logaritmo natural da razão de chances associada a um incremento arbitrário de c unidades em x. O logaritmo natural da razão de chances correspondente a essa variação é dado pela diferença $g(x+c) - g(x) = c\beta_1$, enquanto o *odds ratio* associado é obtido pela exponenciação desse valor: $OR(c) = OR(x+c, x) = e^{c\beta_1}$.

3.2.7 Estratégias de Análise

Para garantir a qualidade e validade dos resultados obtidos, a amostra total foi dividida em dois subconjuntos: um conjunto de treinamento, representando 70% da amostra e um conjunto de teste, representando 30% da amostra total. O conjunto de treinamento foi usado para ajustar o modelo e identificar os padrões presentes nos dados, enquanto o conjunto de teste foi reservado para avaliar a capacidade preditiva do modelo em dados inéditos, minimizando o risco de sobre ajuste (Hair Jr. et al., 2009).

Além disso, o desempenho do modelo foi validado através da análise da Curva ROC (Receiver Operating Characteristic), que permite quantificar a acurácia por meio do cálculo da AUC (Área sob a Curva) (Hanley & McNeil, 1982). Esse método facilita a identificação do limiar óptimo que maximiza a sensibilidade e a especificidade, proporcionando uma avaliação mais detalhada da capacidade discriminatória do modelo. Para interpretação dos valores de AUC, foram adoptados os critérios estabelecidos por Hosmer & Lemeshow (2013):

- AUC ≥ 0.9: Discriminação excelente
- 0.8 ≤ AUC < 0.9: Discriminação boa
- 0.7 ≤ AUC < 0.8: Discriminação aceitável
- 0.6 ≤ AUC < 0.7: Discriminação pobre
- AUC < 0.6: Discriminação inadequada

CAPÍTULO 4

Resultados e Discussão

4.1 Análise Descritiva dos Dados

Foram analisados 9761 indivíduos economicamente activa, definidos como aqueles com 15 anos ou mais, conforme o Instituto Nacional de Estatística. A Figura 4.1 apresenta a distribuição dos indivíduos por sexo. Observa-se que 53.7% dos indivíduos são do sexo feminino, enquanto 46.3% são do sexo masculino.

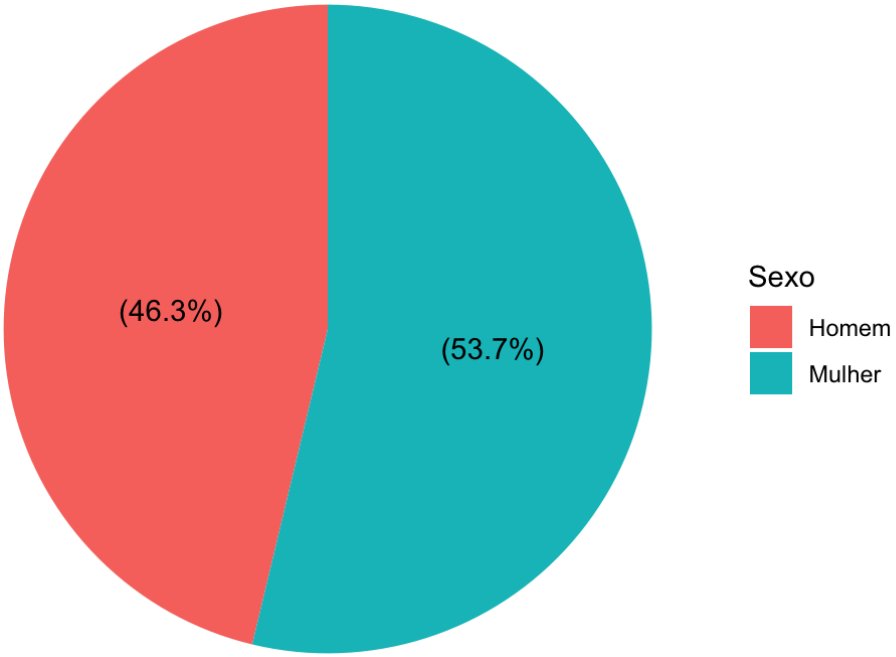

Figura 4.1: Distribuição dos indivíduos por sexo

De acordo com a Figura 4.2, que representa a distribuição da população por idade, observa-se que a maior parte da população (32.4%) está na faixa etária de 15 a 24 anos, caracterizando um perfil populacional relativamente jovem. As demais faixas etárias apresentam uma redução gradual no percentual: 24.3% entre 25 e 34 anos, 18.4% entre 35 e 44 anos, 11.2% entre 45 e 54 anos, e 13.7% acima de 55 anos.

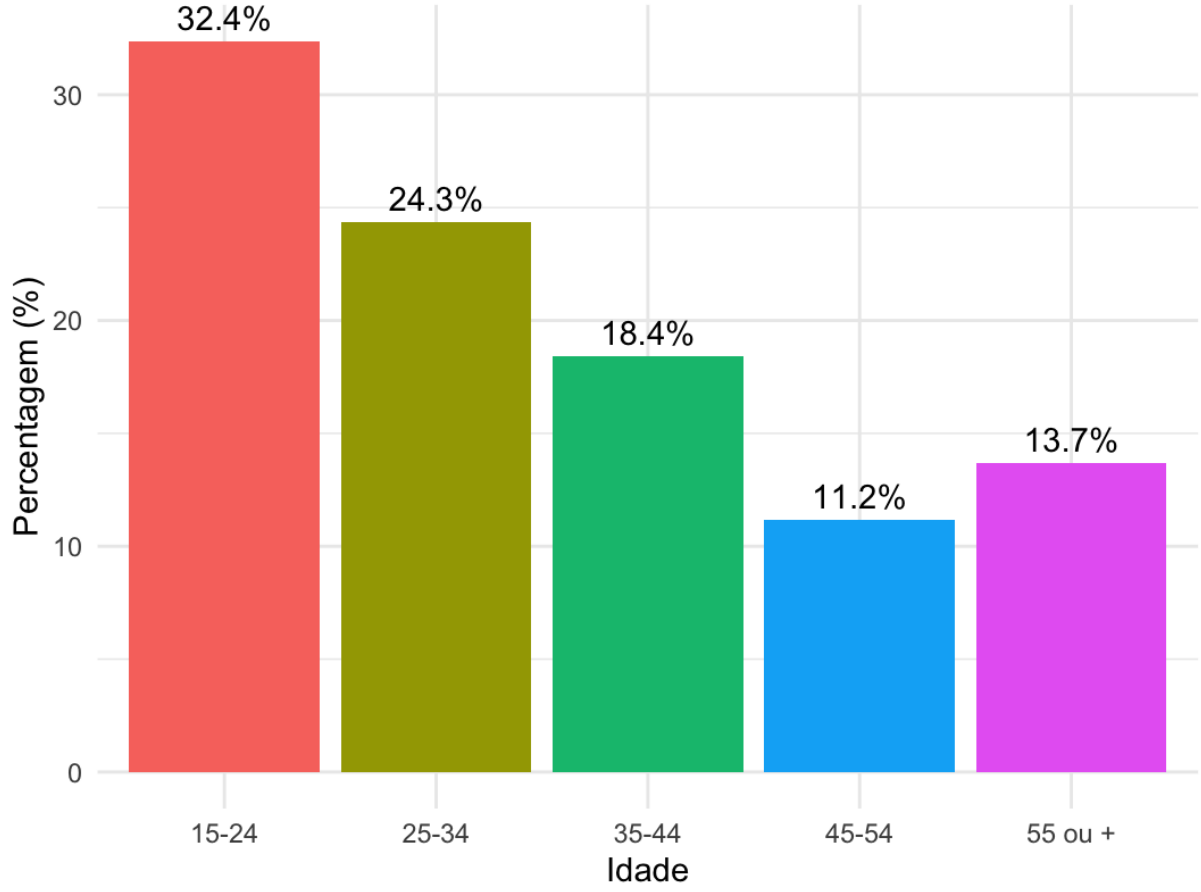

Figura 4.2: Distribuição dos indivíduos por idade

4.2 Associação entre as Variáveis Explicativas e o Desemprego

A Tabela 4.1 mostra a associação entre as variáveis explicativas e o desemprego. Em relação ao género, não existe associação estatisticamente significativa com o desemprego ($p = 0.837$). A proporção de mulheres desempregadas (23.86%) é ligeiramente maior do que a de homens (23.62%). Quanto à idade, existe associação estatisticamente significativa com o desemprego. As faixas etárias intermediárias apresentam as maiores taxas de desemprego, destacando-se o grupo de 25 a 34 anos (32.17%) e de 15 a 24 anos (29.60%). Em contraste, indivíduos com 55 anos ou mais apresentam a menor taxa (8.36%).

O nível de escolaridade demonstra associação estatisticamente significativa com o desemprego. Indivíduos sem escolaridade apresentam taxa de 15.94%, enquanto aqueles com ensino superior registam 11.11%. O estado civil também apresenta associação estatisticamente

significativa, com pessoas casadas apresentando menor taxa de desemprego (20.05%) comparativamente aos não casados (27.12%).

A região de residência influencia significativamente o desemprego, com moradores rurais apresentando uma taxa de desemprego superior (49.57%) aos urbanos (41.64%). Quanto ao grau de parentesco, não chefes de família registam maior desemprego (28.74%) comparativamente aos chefes (14.54%). Surpreendentemente, indivíduos com alguma deficiência apresentam menor taxa de desemprego (10.14%) que aqueles sem deficiência (24.20%).

O tamanho do agregado familiar apresenta uma associação estatisticamente significativa com o desemprego. Observa-se uma tendência de aumento das taxas de desemprego em agregados maiores, com famílias de 9 a 13 membros registando 28.32% e de 14 a 20 membros apresentando 29.09%, comparativamente aos 23.10% observados em agregados menores (1 a 3 membros).

Esses resultados evidenciam a importância de considerar factores demográficos e socioeconómicos na análise do desemprego.

.

Tabela 4.1: Associação entre as variáveis explicativas e o desemprego

Variáveis	Categorias	Desempregado		P-valor
		Não(%)	Sim(%)	
Sexo	Homem	76.38	23.62	**0.837**
	Mulher	76.14	23.86	
Idade	15-24	70.40	29.60	**<0.001**
	25-34	67.83	32.17	
	35-44	80.27	19.73	
	45-54	86.92	13.08	
	55 ou +	91.64	8.36	
Nível de escolaridade	Primário	76.15	23.85	**<0.001**
	Secundário	71.51	28.49	
	Sem escolaridade	84.06	15.94	
	Superior	88.89	11.11	
Casado (a)	Não	72.88	27.12	**<0.001**
	Sim	79.95	20.05	
Região	Rural	50.43	49.57	**<0.001**
	Urbano	58.36	41.64	
Grau de parentesco	Chefe	85.46	14.54	**<0.001**
	Não chefe	71.26	28.74	
Deficiência	Não	75.80	24.20	**<0.001**
	Sim	89.86	10.14	
Tamanho do agregado	1-3	76.90	23.10	**<0.001**
	4-8	76.82	23.18	
	9-13	71.68	28.32	
	14-20	70.91	29.09	

4.3 Modelo de Regressão Logística Ajustado

Os resultados apresentados na Tabela 4.2 mostram a influência das variáveis associadas ao desemprego. São analisados os coeficientes, as constantes, a razão de chance e o p-valor.

Tabela 4.2: Coeficientes Estimados do Modelo de Regressão Logística Binária

	$\widehat{\beta}$	OR	E.P($\widehat{\beta}$)	Wald	P-valor
Constante	-1.540	0.214	0.136	-11.355	**<0.001**
Idade-15-24 (ref)					
25-34	0.385	1.469	0.080	4.820	**<0.001**
35-44	-0.123	0.884	0.101	-1.216	**0.224**
45-54	-0.562	0.570	0.138	-4.071	**<0.001**
55 ou +	-0.993	0.370	0.152	-6.540	**<0.001**
Sexo-*Homem (ref)*					
Mulher	-0.104	0.901	0.064	-1.623	0.105
Tamanho do agregado-*1-3 (ref)*					
4-8	-0.123	0.884	0.076	-1.630	0.103
9-13	-0.001	1.000	0.111	-0.010	0.992
14-20	0.034	1.034	0.390	0.086	0.931
Casado(a)-*Não(ref)*					
Sim	-0.185	0.831	0.069	-2.667	**0.008**
Escolaridade -*Primário(ref)*					
Secundário	0.061	1.063	0.066	0.930	0.352
Sem escolaridade	-0.072	0.931	0.109	-0.656	0.512
Superior	-0.924	0.397	0.176	-5.239	**<0.001**
Região-*Rural(ref)*					
Urbano	0.295	1.344	0.093	3.183	**0.001**
Grau de Parentesco-Chefe*(ref)*					
Não	0.586	1.797	0.088	6.628	**<0.001**

AIC: 6621.9

A probabilidade de um indivíduo estar desempregado é dada por:

$$P_{(\text{Desemprego}=1)} = \frac{e^{Z_i}}{1+e^{Z_i}}$$

Onde:

$$Z = -1.540 + 0.385Idade_{25-34} - 0.562Idade_{45-54} - 0.993Idade_{55+} - 0.185Casado_{\text{Sim}} - 0.924\text{Escolaridade}_{\text{Superior}} + 0.295Região_{\text{Urbana}} + 0.586parentesco_{\text{não chefe}}$$

Idade: Controlando o efeito das demais variáveis, comparado ao grupo de referência (15 a 24 anos), indivíduos na faixa etária de 25 a 34 anos têm 47% mais chance de estar desempregados. A faixa de 35 a 44 anos não é estatisticamente significativa (p = 0,224). Indivíduos de 45 a 54 anos apresentam 43% menos chance de desemprego, enquanto aqueles com 55 anos ou mais têm 63% menos chance de estar desempregados.

Escolaridade: Relativamente ao ensino primário, indivíduos com ensino superior têm 60% menos chance de desemprego. As categorias ensino secundário e sem escolaridade não apresentaram efeitos estatisticamente significativos (p = 0,352 e p = 0,512, respectivamente).

Estado Civil: Indivíduos casados têm 17% menos chance de desemprego em relação aos não casados, controlando o efeito dos demais factores.

Região: Residir em áreas urbanas aumenta em 34% a chance de desemprego comparado a áreas rurais.

Grau de Parentesco: Indivíduos que não são chefes de família têm 80% mais chance de estar desempregados comparativamente aos chefes de família.

Sexo: A variável sexo não apresentou efeito estatisticamente significativo sobre o desemprego (p = 0,105), sugerindo que, controlando o efeito das demais variáveis, não há diferença significativa entre homens e mulheres.

Tamanho do Agregado: As diferentes categorias de tamanho do agregado familiar não demonstraram efeito estatisticamente significativo sobre o desemprego (p > 0,05), indicando que esta variável não é um preditor significativo no modelo ajustado.

4.4 Avaliação do Desempenho do Modelo

O teste de ajustamento de Hosmer e Lemeshow foi aplicado para avaliar o ajuste do modelo. Os resultados estão na Tabela 4.3. O p-valor de 0.1083 indica que não há evidências suficientes para rejeitar a hipótese nula. Assim conclui-se que o modelo se ajusta bem aos dados.

Tabela 4.3: Teste de Hosmer e Lemeshow

Qui-Quadrado	GL	P-valor
10.413	6	0.1083

A análise de multicolinearidade foi realizada para verificar a independência entre as variáveis explicativas do modelo. Os resultados são apresentados na Tabela 4.4, mostrando os valores do Factor de Inflação da Variância (VIF) e a Tolerância para cada variável. Os valores de VIF estão todos abaixo de 10, e as tolerâncias estão acima de 0.1, indicando que não há preocupações significativas de multicolinearidade entre as variáveis do modelo. Isso sugere que as variáveis explicativas são suficientemente independentes umas das outras, permitindo uma interpretação confiável dos coeficientes do modelo.

Tabela 4.4: Multicolinearidade

Variáveis	*Multicolinearidade*	
	VIF	Tolerância
Idade	1.835	0.545
Sexo	1.142	0.876
Tamanho do agregado	1.078	0.928
Casado	1.314	0.761
Nível de educação	1.223	0.818
Região	1.023	0.977
Grau de parentesco	1.498	0.668

A Figura 4.3 ilustra a relação entre a sensibilidade (taxa de verdadeiros positivos) e a especificidade (taxa de verdadeiros negativos) do modelo regressão logística. O valor da AUC (Área Sob a Curva) de 0.7014 indica que o modelo possui uma capacidade discriminativa aceitável.

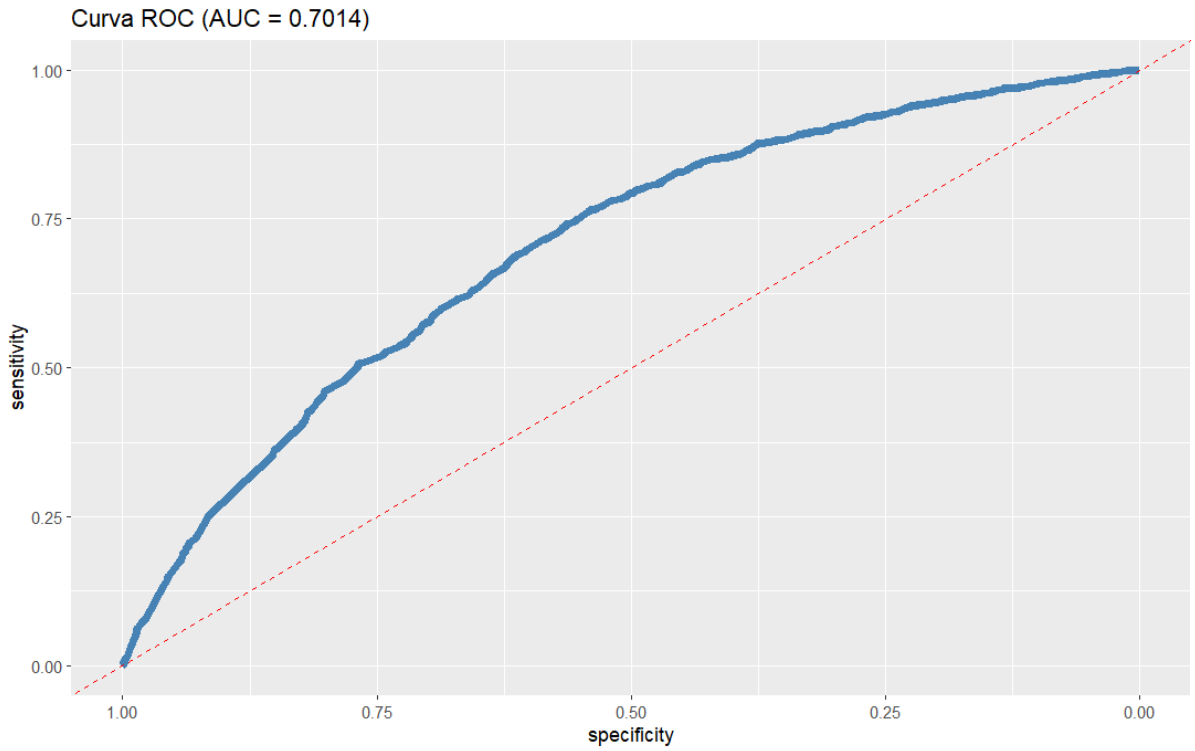

Figura 4.3: Capacidade de Discriminação do Modelo de Regressão Logística

4.5 Discussão de resultados

Esta discussão é estruturada de forma a abordar os principais determinantes do desemprego na cidade e província de Maputo, com base nos dados analisados e nas evidências empíricas encontradas. Os resultados são comparados com a informação da literatura apresentada na secção 2, buscando identificar convergências e divergências, e destacando as particularidades do contexto local.

Os resultados deste estudo indicam que a faixa etária é um factor determinante do desemprego na cidade e província de Maputo. Indivíduos com idades entre 45 e 54 anos têm 43% menos chances de estarem desempregados em comparação aos jovens de 15 a 24 anos, enquanto aqueles com 55 anos ou mais têm uma redução ainda maior, de 63%. Esses achados corroboram com a literatura, como demonstrado por Duarte (2021), que indivíduos mais velhos tem mais chances de empregabilidade em relação aos mais jovens.

O nível de escolaridade surge como um dos factores mais importantes na determinação do desemprego. Indivíduos com ensino superior têm 60% menos chance de desemprego em comparação aos que possuem apenas ensino primário

Esses resultados são consistentes com os achados de Cunha et al. (2011), que destacaram a relação inversa entre nível educacional e desemprego. Indivíduos com maior escolaridade têm maior chances de empregabilidade, especialmente em sectores que demandam competências técnicas e especializadas.

Indivíduos casados têm 17% menos chances de estarem desempregados em comparação aos não casados, controlando o efeito dos demais factores. Este achado é corroborado por Borchers et al. (2022), que identificaram menor probabilidade de desemprego entre pessoas casadas devido à maior pressão por estabilidade financeira e responsabilidades familiares.

Os resultados indicam que indivíduos residentes em áreas urbanas têm 34% mais chances de estarem desempregados em comparação aos que vivem em áreas rurais. Esse achado está alinhado com o estudo de Costa e Cunha (2010), que destacaram maiores taxas de desemprego em áreas urbanas devido à alta densidade populacional e à maior competição por vagas.

Indivíduos que não são chefes de família apresentam 80% mais chances de estarem desempregados em comparação aos chefes de família. Esse achado está em linha com os resultados de Silva et al. (1999), que evidenciaram menores taxas de desemprego entre chefes de família devido à maior responsabilidade financeira e à necessidade de sustentar o núcleo familiar.

CAPÍTULO 5

Conclusões e Recomendações

5.1 Conclusões

O desemprego é um problema socioeconómico, tanto globalmente quanto em Moçambique, com a Cidade e Província de Maputo apresentando taxas superiores à média nacional. O presente estudo teve como objectivo analisar os factores que influenciam a ocorrência do desemprego na Cidade e Província de Maputo. Após a realização do estudo, conclui-se que:

- Dos 9761 indivíduos da amostra, verificou-se que 53.7% eram do sexo feminino e 46.3% do sexo masculino, onde a maior parte eram da faixa etária compreendida entre 15 e 24 anos.
- Indivíduos mais jovens, apresentam maior probabilidade de desemprego. Isso pode ser atribuído à falta de experiência e à competição por vagas em um mercado de trabalho com uma capacidade de absorção limitada.
- O nível de escolaridade mostrou ser um factor crucial. Indivíduos com ensino superior têm significativamente menos chance de estar desempregados, indicando que a qualificação educacional é essencial para melhorar as oportunidades de emprego.
- Pessoas casadas tendem a ter menores taxas de desemprego, Isso pode ser devido à pressão por estabilidade financeira e responsabilidades familiares que incentivam a busca por emprego.
- Residir em áreas urbanas está associado a uma maior chance de desemprego, reflectindo a intensa competição por vagas que caracteriza essas regiões.
- Indivíduos que não são chefes de família têm maior probabilidade de desemprego.
- O modelo de regressão logística se ajustou bem aos dados, com o valor da AUC de 0.7014 indicando que o modelo possui uma capacidade discriminativa aceitável.

5.2 Recomendações

Com base nas conclusões deste trabalho sobre factores que influenciam a ocorrência do desemprego na Cidade e Província de Maputo, as seguintes recomendações são propostas:

- **Realizar Novas Pesquisas**: É crucial realizar novas pesquisas para aprofundar o entendimento dos factores que influenciam a ocorrência do desemprego na região. Pois compreender esses factores é essencial para desenvolver políticas eficazes que possam reduzir as elevadas taxas de desemprego.
- **Incluir Variáveis Adicionais**: Futuras pesquisas devem considerar a inclusão de variáveis adicionais, como factores macroeconómicos, socioeconómicos e de saúde mental, que possam influenciar o desemprego.
- **Utilizar Métodos Estatísticos Avançados**: Aplicar métodos estatísticos mais avançados, como modelos de aprendizado de máquina, pode fornecer uma compreensão mais clara das relações entre as variáveis explicativas e o desemprego. Isso permitirá uma análise mais precisa e preditiva.

5.3 Limitações

O presente trabalho teve algumas limitações, entre elas, a não disponibilidade de informação relativa a importantes factores como naturalidade, experiência profissional, área de formação, raça, acesso ao crédito.

Referências

- Aga, G., Campos, F., Conconi, A., Davies, E., & Geginat, C. (2021). Informal firms in Mozambique: Status and potential (Policy Research Working Paper 9712). World Bank Group.
- Agresti, A. (2013). *Categorical data analysis.* 3rd ed. Wiley.
- Agresti, A. & Finlay, B. (2009). *Statistical Methods for Social Sciences*. 4th ed. Prentice Hall.
- Akaike, H. (1974). A new look at- the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.
- Belfiore, P. (2015). *Estatística aplicada a administração, contabilidade e economia com excel em SPSS. Exemplo de aplicação ao Resolving Credit*. Porto. Vida Económica.
- Borchers, J., Rosalem, L. J. O. S., Leite, T. H., & Araújo, E. (2022). Análise dos determinantes do desemprego e da informalidade juvenil no Brasil (2012-2021). Anais do Encontro Nacional de Economia.
- Brandolini, A., Cipollone, P., & Viviano, E. (2006). Does the ILO definition capture all unemployment? Journal of the European Economic Association, 4(1), 153-179.
- Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.). Springer.
- Cabral, C. S. (2013). Aplicação do modelo de regressão logística num estudo de mercado [Dissertação de mestrado, Universidade de Lisboa]. Repositório da ULisboa.
- Companhia de Planejamento do Distrito Federal. (2021). Pessoas com deficiência (PcD) e mercado de trabalho no Distrito Federal.
- Cordeiro, G. M., & Demétrio, C. G. B. (2008). *Modelos Lineares Generalizados e Extensões*. Piracicaba: USP.
- Costa, J. S., & Cunha, M. S. (2010). Determinantes do desemprego no Brasil no período de 1981 a 2005.
- Cunha, D. A., Araújo, A. A., & Lima, J. E. (2011). Determinantes do desemprego e inatividade de jovens no Brasil metropolitano. *Revista de Economia e Agronegócio*, 9(3), 369-392.

- Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models* (4th ed.). Chapman and Hall/CRC.
- Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). Wiley.
- Duarte, L. B. (2021). Utilização do modelo logit para analisar os determinantes do desemprego.
- Fávero, L. P., & Belfiore, P. (2017). Manual de análise de dados: Estatística e modelagem multivariada com Excel, SPSS e Stata. Elsevier.
- Ferrão, J., Bell, V., Cardoso, L. A., & Fernandes, T. (2018). Agriculture and Food Security in Mozambique. Journal of Food, Nutrition and Agriculture, 1(1), 7-11. https://doi.org/10.21839/jfna.2018.v1i1.121
- Figueira, C. V. Modelos de regressão logística. (2006). Universidade Federal do Rio Grande de Sul. Porto Alegre. Dissertação de mestrado
- Fundação Fé e Cooperação. (2023). Estudo sobre o mercado de trabalho na cidade de Maputo. FEC - Fundação Fé e Cooperação.
- Gonzalez, L. A. (2018). Regressão Logística e suas aplicações. Universidade Federal de Maranhão. Dissertação de mestrado.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Multivariate data analysis* (6th ed.). Porto Alegre: Bookman.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Hussmanns, R. (2007). Measurement of employment, unemployment and underemployment – Current international standards and issues in their application. Bulletin of Labour Statistics, 1, XI-XXIX.
- Instituto Nacional de Estatística. (2019). IV Recenseamento Geral da População e Habitação 2017: Resultados Definitivos - Moçambique. INE.
- Instituto Nacional de Estatística. (2023). Inquérito sobre Orçamento Familiar - IOF 2022. Maputo, Moçambique: INE.
- International Finance Corporation & L.E.K. Consulting. (2019). Digital Skills in Sub-Saharan Africa: Spotlight on Ghana. International Finance Corporation

- International Labour Organization. (2013). Resolution concerning statistics of work, employment and labour underutilization. 19th International Conference of Labour Statisticians. Geneva: ILO.
- International Labour Organization. (2018). Measuring unemployment and the potential labour force. ILO Department of Statistics. Retrieved from https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@stat/documents/publication/wcms_627878.pdf
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.
- Kapur, R. (2022). Types of unemployment: Unfavourable in leading to up-gradation of living conditions of individuals. Recuperado de https://www.researchgate.net/publication/363054513_Types_of_Unemployment_Unfavourable_in_leading_to_Up-gradation_of_Living_Conditions_of_Individuals
- Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4419-1742-3
- Lindsey, J. K. (1997). Applying Generalized Linear Models. Springer.
- Maroco, J. (2007). *Análise estatística Com Utilização do SPSS*, 3ª Edição. Lisboa: Edição Silabo.
- McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models (2nd ed.). Chapman and Hall.
- Meyer, A. N. (1980). Análise da qualidade de vida no trabalho utilizando um modelo de regressão logística [Dissertação de mestrado]. Universidade Tecnológica Federal do Paraná.
- Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). Applied multivariate research: Design and interpretation. Sage Publications.
- Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3), 253-267. https://doi.org/10.1080/09720502.2010.10700699
- Ministério do Gênero, Criança e Ação Social. (2022). Gender Equality Profile Mozambique.
- Msigwa, R., & Kipesha, E. F. (2013). Determinants of Youth unemployment in Developing Countries: Evidences from Tanzania. *Journal of Economics and Sustainable Development*, 4(14), 67-76.

- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.
- Oliveira, P. R., Scorzafave, L. G., & Pazello, E. T. (2009). Desemprego e inatividade nas metrópoles brasileiras: as diferenças entre homens e mulheres. *Nova Economia, 19*(2), 291-324.
- Organização Internacional do Trabalho. (2020). World Employment and Social Outlook: Trends 2020. OIT.
- Organização Internacional do Trabalho. (2021). World Employment and Social Outlook 2021: The role of digital labour platforms in transforming the world of work. OIT.
- Paula, G. A. (2010). *Modelos de Regressão com Apoio Computacional*. IME-USP.
- Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14. https://doi.org/10.1080/00220670209598786
- Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2ª ed.). Sage Publications.
- Silva, D. B. N., Melo, D. L. B., & Lima, J. M. (1999). Determinantes do desemprego em comunidades de baixa renda da cidade do Rio de Janeiro. VI Encontro Nacional de Estudos do Trabalho – ABET.
- Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia Medica, 24(1), 12–18. https://doi.org/10.11613/BM.2014.003
- Statistics Canada. (2015). Measuring Employment and Unemployment in Canada and the United States – A comparison. Labour Statistics: Technical Papers, 75-005-M. Retrieved from https://www150.statcan.gc.ca/n1/pub/75-005-m/75-005-m2015002-eng.htm
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.
- Turkman, M. A. A., & Silva, G. L. (2000). *Modelos Lineares Generalizados: da teoria à prática*. Lisboa: Sociedade Portuguesa de Estatística.
- United Nations Mozambique. (2021). Common Country Analysis: Mozambique. UN. Retrieved from https://minio.uninfo.org/uninfo-production-main/1d61d26a-fd42-4733-9cba-908aa48fd272_Final_CCA_Mozambique_-_August_2021.pdf
- UN Women. (2024). Facts and figures: Economic empowerment. https://www.unwomen.org/en/what-we-do/economic-empowerment/facts-and-figures

- World Bank. (2018). Mozambique Jobs Diagnostic (Report No. 129408). World Bank Group.https://documents.worldbank.org/curated/en/655951534181476346/pdf/Mozambique-Jobs-Diagnostic-Volume-1-Analytics.pdf
- World Economic Forum. (2022). Why Africa's youth hold the key to its development potential. https://www.weforum.org/agenda/2022/09/why-africa-youth-key-development-potential
- World Bank. (2023). Unemployment, youth total (% of total labor force ages 15-24) - Sub-Saharan Africa. World Bank Data. https://data.worldbank.org/indicator/SL.UEM.1524.ZS
- Yánez Contreras, M., & Cano Hernández, K. (2011). Determinantes del desemprego: Una revisión de la literatura. *Revista Panorama Económico*, 19, 135-148.
- Zurrón Ocio, D. (1995). O emprego na teoria econômica. Relatório de Pesquisa Nº 11/1995. Escola de Administração de Empresas de São Paulo, Fundação Getulio Vargas