Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Modelação Estatística e Análise de Dados

Relatório de Estágio

Análise de indicadores sobre o acesso às respostas sociais de apoio à "Família e Comunidade": Modelos de Análise de regressão e correlação entre as variáveis

Catarina Wengorovius Viana de Sousa

Orientador(es) | Dulce Gamito Pereira

Antonieta do Rosário Pinto Sebastião Rodrigues Ministro

Évora 2024

Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Modelação Estatística e Análise de Dados

Relatório de Estágio

Análise de indicadores sobre o acesso às respostas sociais de apoio à "Família e Comunidade": Modelos de Análise de regressão e correlação entre as variáveis

Catarina Wengorovius Viana de Sousa

Orientador(es) | Dulce Gamito Pereira

Antonieta do Rosário Pinto Sebastião Rodrigues Ministro

Évora 2024

O relatório de estágio foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor da Escola de Ciências e Tecnologia:

Presidente	|	Lígia Henriques-Rodrigues (Universidade de Évora)
Vogais	|	Dulce Gamito Pereira (Universidade de Évora) (Orientador) Luís M. Grilo (Universidade de Évora)

Évora 2024

Agradecimentos

Os meus agradecimentos são dirigidos a todos os que direta ou indiretamente tornaram possível a concretização deste trabalho.

Em especial, agradeço à minha família, grande, barulhenta e sempre presente.

À Professora Dulce Pereira, pela orientação, partilha de conhecimentos, e toda a atenção, que muito contribuíram para a realização deste relatório.

À Dra Antonieta pela abertura e pela orientação nas ideias, no trajeto e na concretização do tema, pela disponibilidade, transmitindo-me a sua confiança.

A toda a equipa do Gabinete de Estratégia e Planeamento, que me receberam e acolheram com o maior cuidado desde o primeiro dia.

Aos meus amigos, que tiveram sempre palavras simpáticas e um abraço para oferecer.

Análise de indicadores sobre o acesso a equipamentos sociais e a igualdade de género: Modelos de Análise de Regressão e Correlação, e Modelos Estatísticos de Previsão

Resumo

Dada a importância de se obter um maior conhecimento sobre a pobreza em Portugal, este trabalho tem como principal objetivo analisar os possíveis indicadores/variáveis com influência/impacto na proporção da população residente em risco de pobreza ou exclusão social em Portugal. Para além disso, analisa-se se, e em que medida, é que as respostas sociais de apoio à Família e Comunidade mostram ter impacto e a sua correlação com a evolução da pobreza em Portugal.

Consideram-se dados reais dos indicadores de pobreza e desigualdade entre 2002 e 2022, sendo a sua análise feita a nível nacional.

Efetua-se uma investigação básica de tipo descritivo para uma caracterização das diferentes variáveis e investigam-se as correlações entre elas através do coeficiente de Pearson.

Com base na análise de regressão identificam-se as variáveis que melhor permitem compreender o fenómeno da pobreza. Para tal, o trabalho encontra-se dividido em quatro modelos de regressão linear múltipla distintos.

Palavras-Chave:

Análise de Regressão e Correlação. Modelos de Regressão Linear Múltipla. Pobreza e desigualdades sociais. Respostas Sociais. Família e Comunidade.

Analysis of indicators on access to social facilities and gender equality: Regression and Correlation Analysis Models, and Statistical Forecasting Models

Abstract

Given the great need to obtain greater knowledge about poverty in Portugal, this work's main objective is to analyze the indicators/variables that have the greatest influence/impact on the proportion of the resident population at risk of poverty or social exclusion in Portugal. Furthermore, it is analyzed whether, and to what extent, social facilities supporting the Family and Community have an impact and their correlation with the evolution of poverty in Portugal.

Real data on poverty indicators between 2002 and 2022 are considered, and the analysis of these data is carried out at national level.

A basic descriptive investigation was carried out to characterize the different variables and the correlations between them were investigated using the Pearson coefficient.

Based on regression analysis, the variables that best allow us to understand the phenomenon of poverty were identified. To this end, the work is divided into four distinct multiple linear regression models.

Keywords: Regression and Correlation Analysis. Multiple Linear Regression Models. Poverty and social inequalities. Social Responses. Family and Community.

Índice

1. Introdução 1

2. Apresentação da Instituição de acolhimento 2

3. Enquadramento geral 3

3.1 Conceitos 3

3.1.1 Pobreza 3

3.1.2 População em risco de pobreza ou exclusão social 4

3.1.3 Programas de ajuda material e medidas de inclusão social 5

3.1.4 Rede de Serviços e Equipamentos Sociais 5

3.1.5 Equipamentos existentes para o grupo-alvo Família e Comunidade 6

4. Revisão da Literatura 7

5. Utilização da Regressão Linear Múltipla 8

5.1 Coeficiente de correlação de Pearson 10

5.2 Coeficiente de determinação 11

5.3 Teste à significância global do modelo 12

5.4 Testes de significância individual 12

5.5 Pressupostos da Regressão Linear Múltipla 13

5.5.1 Os resíduos seguem uma distribuição normal 13

5.5.2 Têm valor médio nulo 14

5.5.3 Variância constante 14

5.5.4 Independência dos resíduos 15

5.5.5 Ausência de multicolinearidade 15

5.6 Existência de *outliers* ou observações influentes 15

5.7 Métodos estatísticos de seleção de variáveis 16

5.8 Previsão 17

6. Modelação Estatística 19

6.1 Primeiro Modelo 20

6.1.1 Estatística descritiva da variável dependente do modelo 23

6.1.2 Análise de correlação 26

6.1.3 Modelação 27

6.2 Segundo Modelo 38

6.2.1 Estatística descritiva da variável dependente 39

6.2.2 Análise de correlação 40

6.2.3 Modelação 41

6.3 Terceiro Modelo 51

6.3.1 Estatística descritiva das variáveis RSES família e comunidade 52

6.3.2 Modelação 54

6.4 Quarto Modelo 63

6.4.1 Modelação 63

Conclusão 71

Bibliografia 73

Índice de Figuras

Figura 1 - Taxa de utilização das respostas sociais dirigidas à Família e Comunidade (%) 6
Figura 2 - Medidas de dispersão da variável PPRPES 25
Figura 3 - Valores VIF do Modelo inicial 28
Figura 4 - Valores VIF do Modeloa 28
Figura 5 - Modelob 29
Figura 6 - Modelobfinal 29
Figura 7 - Modelod final 31
Figura 8 - Normalidade dos Resíduos 33
Figura 9 - Resíduos vs Valores estimados 34
Figura 10 - Distância de Cook 35
Figura 11 - Valores de Leverage 35
Figura 12 - Medidas de dispersão da variável TIP 40
Figura 13 - Gráfico de Correlação 41
Figura 14 - Modelo 2 - Modelo inicial 42
Figura 15 - Modelo 2 - Modelo final 43
Figura 16 - Valores de Leverage e Distância de Cook 48
Figura 17 - Nº de respostas Sociais, Portugal - 2021 52
Figura 18 - Capacidade e Frequência das respostas sociais, Portugal - 2021 53
Figura 19 - "Melhor modelo" pelo método estatístico "Backward" 55
Figura 20 - Modelo 3 - Modelo final 55
Figura 21 - Q-Q Plot Normalidade 57
Figura 22 - "Melhor modelo" pelo método estatístico "Stepwise" 63
Figura 23 - Modelo 4 - Modelo final 64

Lista de Tabelas

Tabela 1 – Interpretação do coeficiente de correlação _____ 10
Tabela 2 – Variáveis do modelo _____ 20
Tabela 3 – Coeficientes de Correlação de Pearson _____ 25
Tabela 4 – Previsão da variável dependente PPRPES _____ 36
Tabela 5 – Cenário otimista _____ 37
Tabela 6 – Variáveis do modelo _____ 38
Tabela 7 – Previsão da variável dependente TIP _____ 48
Tabela 8 – Cenário _____ 48
Tabela 9 – Dois cenários _____ 50
Tabela 10 – Variáveis do modelo _____ 51
Tabela 11 – Análise das Respostas Sociais _____ 51
Tabela 12 – Previsão da variável dependente PPRPES _____ 60
Tabela 13 – Cenário 1 _____ 61
Tabela 14 – Cenário 2 _____ 61
Tabela 15 – Cenário 3 _____ 62
Tabela 16 – Previsão da variável dependente PPRPES _____ 67
Tabela 17 – Cenário _____ 68

1. Introdução

No presente relatório pretende-se apresentar um estudo que analisa elementos relativos à evolução do fenómeno da pobreza e das desigualdades em Portugal, e alguns dos indicadores/variáveis que têm maior influência/impacto na proporção da população residente em risco de pobreza ou exclusão social. Para além disso, analisa-se se, e em que medida, é que as respostas sociais de apoio à Família e Comunidade mostram ter impacto e a sua correlação com a evolução da pobreza em Portugal.

Para tal foi utilizada a Análise de Regressão, com a qual foram aplicados modelos de Regressão Linear Múltipla, analisados através do *software* estatístico R.

O objetivo deste estudo encontra-se, assim, dividido em duas partes:

1º Análise da evolução da pobreza e desigualdades em Portugal.

2º Aferição da ligação da evolução da pobreza com as respostas sociais existentes.

Considerando a amplitude do tema e a quantidade de variáveis que poderiam ser atendidas para conseguir obter resultados válidos, centramo-nos em quatro modelos de Regressão Linear Múltipla, explicitando o seu modo de funcionamento, tendo-se optado neste estudo pela abordagem de pobreza multidimensional.

Compreendido o conceito e a evolução dos vários indicadores, o estudo focou-se na perceção dos efeitos da sua correlação com as respostas sociais existentes em Portugal, designadamente as dirigidas ao grupo-alvo "Família e Comunidade", as quais englobam respostas sociais que têm como objetivo o apoio a pessoas e famílias que se encontrem em situação de vulnerabilidade, exclusão ou de marginalização social.

Na secção 2 apresenta-se a Instituição de acolhimento do estágio. A secção 3 é dedicada ao enquadramento geral da problemática do fenómeno da pobreza, nomeadamente aos conceitos de pobreza, população em risco de pobreza ou exclusão social, apresentação dos programas de ajuda material e medidas de inclusão social em vigor, caracterização da Rede de Serviços e Equipamentos Sociais disponíveis e indicação das respostas existentes para o grupo –alvo "Família e Comunidade".

Na secção 4 encontra-se uma revisão da literatura, expõem-se estudos sobre esta temática, realizados noutros países, que mostram a aplicação da Análise de Regressão e

a construção de modelos de Regressão Linear Múltipla, onde, em alguns deles, foi utilizado também o apoio do *software* estatístico R.

Na secção 5, apresenta-se o modelo de Regressão Linear Múltipla e, enquadrado na temática, dedicou-se uma subsecção aos seguintes tópicos: pressupostos do modelo, pesquisa de observações influentes, métodos de seleção de variáveis e por fim à previsão.

Na secção 6, apresentam-se os estudos realizados, aplicados a dados reais, que dizem respeito aos indicadores de pobreza em Portugal e ao impacto das respostas sociais na mesma. Recorrendo ao uso do *software* estatístico R, exemplificando algumas técnicas descritas no trabalho. Por último, na secção 7 é feita a suma das conclusões apresentadas.[1]

2. Apresentação da Instituição de acolhimento

A escolha da realização de um Estágio Curricular teve em vista a conciliação dos conhecimentos adquiridos durante a minha Licenciatura em Ciência Política e Relações Internacionais com as matérias e ferramentas adquiridas no Mestrado de Modelação Estatística e Análise de Dados, procurando potenciar ambos no exercício de uma atividade concreta.

O Gabinete de Estratégia e Planeamento (GEP) do Ministério do Trabalho, Solidariedade e Segurança Social (MTSSS) foi a instituição escolhida para a realização deste Estágio Curricular. O GEP-MTSSS é um serviço central da administração direta do Estado, dotado de autonomia administrativa (Decreto-Lei n.º 14/2015, de 26 de janeiro). Tem como atribuições, entre outras, a promoção e realização de investigação e estudos prospetivos que contribuam para a definição e estruturação das estratégias, políticas, prioridades e objetivos do MTSSS. Neste âmbito, o GEP-MTSSS é a entidade responsável pelo desenvolvimento e atualização da Carta Social. A Carta Social constitui-se como uma ferramenta essencial ao estudo da dinâmica da Rede de Serviços e Equipamentos Sociais (RSES) e "(…) apresenta-se como um instrumento de informação privilegiado de caraterização e análise, essencial para o processo de conceção e adequação das

[1] As conclusões e opiniões expressas no trabalho apresentado são da exclusiva responsabilidade da autora e não podem ser imputadas (ou atribuídas) a terceiros, designadamente ao GEP ou seus colaboradores.

políticas sociais, para o apoio ao planeamento territorial e à preparação da tomada de decisão, afirmando-se também como meio fundamental na linha de informação ao cidadão."[2]

3. Enquadramento geral

3.1 Conceitos

Nas subsecções seguintes são apresentados alguns conceitos fundamentais para a compreensão da problemática tratada neste estudo.

3.1.1 Pobreza

A pobreza é um fenómeno complexo e multidimensional que não pode ser plenamente compreendido com uma única definição. Responsáveis políticos e investigadores utilizam uma série de indicadores para medir a pobreza e acompanhar a sua evolução ao longo do tempo, com o objetivo de conceber intervenções eficazes para o seu combate e para reduzir as desigualdades.

A pobreza é normalmente medida em termos de rendimento, embora também se tenha em conta outros fatores como a saúde, a educação e a exclusão social. Comecemos por fazer uma breve apresentação dos dois tipos/conceitos de pobreza cuja distinção remonta pelo menos ao final do século XIX. Ao falarmos de pobreza absoluta estamos a definir situações de privação severa e extrema em que a ausência de condições financeiras não permite que um cidadão ou família atenda às suas necessidades básicas para a sobrevivência, como alimentação, abrigo e vestuário. Já a pobreza relativa é definida como situações em que o rendimento de um cidadão ou família é inferior a 60%[3] do rendimento mediano do país, o que significa que os cidadãos que vivem abaixo deste limiar fixado têm um padrão de vida significativamente mais baixo do que a média da população.[4]

Podemos ainda referir uma outra abordagem ao conceito de pobreza: a pobreza multidimensional, que resultou de um processo de evolução conceitual ao longo das

[2] Carta Social – Rede de Serviços e Equipamentos Sociais, Relatório 2020

[3] Ainda que possam ser disponibilizadas outras percentagens na literatura.

[4] Conceitos disponíveis no Portal do INE: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE

últimas décadas, e veio a ser elaborado e promovido por autores como Amartya Sen[5], Sabina Alkire e James Foster [6]. Este conceito insere-se numa abordagem mais ampla e abrangente para calcular e compreender a pobreza, que em vez de se centrar apenas no rendimento e consumo dos cidadãos, como é o caso da abordagem unidimensional de pobreza, compreende outras dimensões que afetam o bem-estar e o desenvolvimento humano, tais como: participação no mercado de trabalho; privação material; privação social; saúde; e habitação (Alves, N., 2022).[7] As variáveis que integram o estudo que é apresentado neste relatório foram selecionadas dentro da abordagem multidimensional referida.

3.1.2 População em risco de pobreza ou exclusão social

Conforme as definições apresentadas acima podemos caracterizar a proporção da população residente em risco de pobreza ou exclusão social[8] como indivíduos que enfrentam situações de privação material, social e/ou cultural severa que os colocam num contexto de vulnerabilidade em relação ao resto da sociedade.[9]

Conforme a definição dada pelo INE: "Indivíduos em risco de pobreza ou em situação de privação material e social severa ou a viver em agregados com intensidade laboral per capita muito reduzida." Sendo que, por "privação material e social severa" entende-se: Condição do agregado doméstico privado no qual se verifica a carência forçada de pelo menos quatro dos nove itens considerados necessários para boas condições de vida, devido a dificuldades económicas - estes serão enumerados mais à

---

[5] Amartya Sen é um Economista, vencedor do Prémio Nobel de Economia em 1998, sendo que foi um dos idealizadores do Índice de Desenvolvimento Humano. É reconhecido internacionalmente pela sua dedicação ao combate à pobreza através de soluções concretas e estratégias complexas.

[6] Sabina Alkire e James Foster são os autores responsáveis pelo desenvolvimento do Índice de Pobreza Multidimensional.

[7] Alves, N. 2022. "Um indicador de pobreza multidimensional para Portugal". *Revista de Estudos Económicos*, Vol. VIII, Nº4, páginas 30-54. Disponível em: https://www.bportugal.pt)

8 Por convenção, este indicador é referenciado ao ano do inquérito. O indicador "População residente em risco de pobreza ou exclusão social" combina dois indicadores construídos com base em informação relativa ao ano de referência do rendimento (Taxa de risco de pobreza após transferências sociais e Intensidade laboral per capita muito reduzida) com um indicador com informação relativa ao ano do inquérito (Taxa de privação material severa). – Portal INE

[9] INE. (2020). Inquérito às Condições de Vida e Rendimento - 2022. Lisboa: INE. Disponível em: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE

frente na explicação das variáveis independentes selecionadas. E com "intensidade laboral per capita muito reduzida" entende-se: "Proporção de indivíduos com menos de 60 anos que, no período de referência do rendimento, viviam em agregados familiares cujos adultos entre os 18 e os 59 anos (excluindo estudantes) trabalharam em média menos de 20% do tempo de trabalho potencial." [10]

A identificação deste grupo é importante para orientar políticas públicas que possam combater os fatores que poderão conduzir a situações de privação, e garantir que todos os cidadãos têm acesso a condições de vida dignas e à participação na sociedade.

3.1.3 Programas de ajuda material e medidas de inclusão social

Para que seja possível combater a realidade da pobreza não basta a ajuda em distribuição de bens materiais de necessidade básica, são necessários apoios e medidas de inclusão social. Foi nesta premissa que se centrou o Fundo de Auxílio Europeu às Pessoas mais Carenciadas (FEAD), em vigor entre 2014 e 2020, período este que se encaixa no período dos dados deste estudo. Como salienta a Comissão Europeia: "A ajuda material deve ser complementada por medidas de inclusão social (nomeadamente, orientação e apoio) para ajudar as pessoas a sair da situação de pobreza em que se encontram"[11].

Atualmente, o FEAD fundiu-se com o *European Social Fund* (ESF), incluindo o *Youth Employment Initiative* (YEI) e o *EU Programme for Employment and Social Innovation* (EaSI), denominados enquanto conjunto por ESF+, que estará em vigor entre 2021 e 2027.

É na execução deste propósito de complementar os programas de ajuda material com medidas de orientação e apoio que se integra a Rede de Serviços e Equipamentos Sociais.

3.1.4 Rede de Serviços e Equipamentos Sociais

Posto isto, ao longo deste Estágio, neste trabalho pretendeu-se estudar e observar

[10] INE. Disponível em: https://smi.ine.pt/ .

[11] Comissão Europeia. Disponível em: https://ec.europa.eu/social/main

a evolução de algumas das respostas sociais integrantes da Rede de Serviços e Equipamentos Sociais (RSES)[12], que oferecem precisamente esta orientação e apoio, referidas no ponto 3.1.3., direcionadas para os grupos mais vulneráveis. A RSES tem um papel determinante no combate às situações de pobreza, assim como na promoção da inclusão social/no combate à exclusão social, e na garantia de igualdade de oportunidades para todos os cidadãos.

O Gabinete de Estratégia e Planeamento (GEP) do Ministério do Trabalho, Solidariedade e Segurança Social é responsável por estudar a evolução da Rede de Serviços e Equipamentos Sociais (RSES), através da Carta Social – instrumento de informação privilegiado de caracterização e análise, para o apoio ao planeamento territorial e à preparação da tomada de decisão.

As respostas sociais, isto é, atividades e/ou serviços desenvolvidos em equipamentos, que compõem a RSES dirigem-se a diferentes públicos-alvo, nomeadamente: Crianças e Jovens; Pessoas com Deficiência ou Incapacidade; Pessoas Idosas; Família e Comunidade. O foco deste estudo irá incidir nas respostas sociais que visam o apoio à Família e Comunidade.

3.1.5 Equipamentos existentes para o grupo-alvo Família e Comunidade

As respostas sociais dirigidas à Família e Comunidade têm como objetivo o apoio a pessoas e famílias que se encontrem em situação de vulnerabilidade, exclusão ou de marginalização social, quer através do apoio e acompanhamento social, quer através da minimização de situações de carência, podendo compreender um conjunto de ações integradas com vista à sua inserção social. (Carta Social - Rede de Serviços e Equipamentos, Relatório 2021).

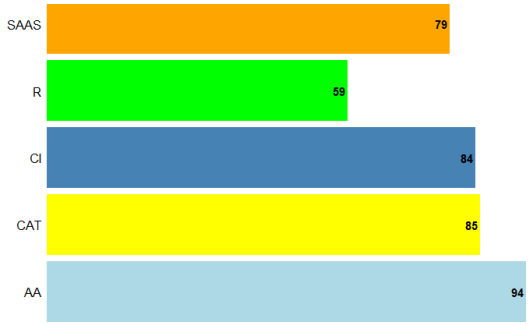

Figura 1 - Taxa de utilização das respostas sociais dirigidas à Família e Comunidade (%)

[12] Carta Social – Rede de Serviços e Equipamentos - Relatório 2021 - https://www.cartasocial.pt/relatorios

O Serviço de Atendimento e Acompanhamento Social (SAAS), a Cantina Social (CS), a Comunidade de Inserção (CI), o Centro de Alojamento Temporário (CAT) e a Ajuda Alimentar (AA) constituem algumas das respostas dirigidas a este grupo-alvo.

O gráfico representado acima (Figura 1) traduz a taxa de utilização em cada uma das respostas sociais em relação à sua capacidade. Sendo que para os dados mais atuais, aos quais tivemos acesso, referentes ao ano de 2021, todas as respostas se encontram com taxas de utilização aproximadamente entre os 80% e os 95%, com exceção do Refeitório/Cantina Social que se encontra com uma taxa de utilização de 59%.

4. Revisão da Literatura

Foram já desenvolvidos diversos estudos semelhantes nesta área da análise da pobreza em vários países e onde foi escolhida a Regressão Linear Múltipla (RLM) como metodologia e técnica estatística.

"The Impact of Socioeconomic and Demographic Variables on Poverty: A Village Study" de Imran Sharif Chaudhry, Shahnawaz Malik e Abo ul Hassan (Chaudhry, Malik, and Hassan, 2009), é um exemplo de um estudo realizado em diferentes cidades do Paquistão, onde os autores analisam a influência de um conjunto de variáveis socioeconómicas e demográficas dos agregados familiares na pobreza (variáveis independentes), através de um modelo de RLM, apresentado por uma equação cujo poder de explicação, medido pelo Coeficiente de Determinação (R^2), é significativamente elevado. Os autores explicam quais as variáveis que se mostraram mais significativas e, desta forma, comprovam que as variáveis socioeconómicas e, em especial, as demográficas têm um impacto significativo no rendimento dos agregados familiares, bem como na redução da incidência da pobreza no Paquistão no geral. Embora as conclusões deste estudo sejam muito interessantes, não foi possível extrair destas matéria que releve para o estudo em análise, visto que os contextos em Portugal e no Paquistão são muito diferentes e as próprias variáveis explicativas acabam por ser muito dissemelhantes, com exceção do índice de dependência total, incluídos em ambos os estudos. Contudo, a leitura e compreensão da estrutura do estudo revelou-se muito útil para a organização deste relatório por permitir ver a aplicação prática desta metodologia.

"Analysis of Effect of GRDP (Gross Regional Domestic Product) Per Capita, Inequality Distribution Income, Unemployment and Human Development Index on Poverty" de Murbanto Sinaga (Sinaga, 2020) é outro exemplo de um estudo que analisa a pobreza, com um olhar economicista, utilizando a RLM como ferramenta estatística. Este foi realizado na Indonésia e teve como variáveis independentes as quatro definidas no título do estudo: produto interno bruto per capita; desigualdade na distribuição dos rendimentos; desemprego; e índice de desenvolvimento humano. Os autores apresentaram a análise de resíduos de forma a validarem todos os pressupostos do modelo de regressão linear múltipla, e escreveram a equação ajustada tendo interpretado os coeficientes ajustados, sendo que neste relatório tentarei desenvolver uma análise semelhante. Em suma, uma das conclusões que retiraram foi a de que apenas uma das variáveis explicativas mostrou ter um efeito estatisticamente significativo na variável dependente pobreza, esta foi a variável desemprego. Este relatório terá, por sua vez, também como um dos objetivos a análise de qual ou quais variáveis mostram ter efeitos mais significativos.

"Mapping Regional Vulnerability to Energy Poverty in Poland" de Lilia Karpinska, Sławomir Smiech, João Pedro Gouveia e Pedro Palma (Karpinska et al., 2021) foi um outro estudo, muito interessante, onde investigadores do Departamento de Estatística da Universidade de Economia de Cracow na Polónia e investigadores da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa se juntaram para analisar mais especificamente a pobreza energética das famílias no território polaco e avaliar a vulnerabilidade regional. Para tal utilizaram como uma das ferramentas estatísticas a criação de modelos de RLM que incluem variáveis que refletem a eficiência energética da habitação e as necessidades energéticas das famílias, e realizaram a análise estatística no *software* R com bases de dados nacionais, *software* que também foi o utilizado para a realização do presente relatório.

5. Utilização da Regressão Linear Múltipla

"Quando se estuda, com base em dados, um determinado fenómeno de natureza social, (…) com o objetivo de descrever, explicar ou prever o seu comportamento, procura-se conceber, ainda que de forma aproximada ou simplificada, o mecanismo subjacente ao fenómeno observável. Este mecanismo é designado habitualmente por

modelo teórico.” (Murteira et al., 2015).

Um dos objetivos fundamentais da ciência é investigar a relação estatística entre fenómenos de forma a melhorar o conhecimento da realidade e a prever a evolução dos fenómenos envolvidos nessa relação. (Pereira, D. (2022) *Regressão Linear e Multilinear* [PowerPoint slides]).

Como forma de descobrir e medir as relações estatísticas que não são definidas nem como relações exatas nem como relações de independência optou-se pela Análise de Regressão, mais precisamente, a construção de modelos de Regressão Linear Múltipla.

A Regressão Linear Múltipla (RLM) é um instrumento estatístico utilizado para analisar a relação entre uma variável dependente e várias variáveis independentes. A RLM é amplamente utilizada em diferentes estudos sobre a pobreza, em vários países, e permite identificar os fatores que contribuem para explicar este fenómeno e a desigualdade num país.

Neste estudo, ajustou-se um modelo de Regressão Linear Múltipla que pode ser resumidamente expresso na seguinte equação matemática:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \,, \quad i = 1, \dots, \mathrm{n}$$

onde,

Y = variável dependente

X_k = variáveis independentes

β_0 = parâmetro, ordenada na origem

β_k = parâmetros, coeficientes da regressão

ε_i = erros ou resíduos aleatórios

Para estimar os coeficientes da regressão é utilizado o método dos mínimos quadrados, que minimiza a soma dos quadrados dos desvios das observações em relação à equação de regressão.

A equação ajustada é dada por:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}, \quad i = 1, \dots, n$$

Sendo os resíduos de estimação $\varepsilon_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$

Em muitas situações, o modelo não é linear (relativamente aos parâmetros), mas mediante uma transformação da variável *Y* consegue obter-se uma relação representada como indica a fórmula acima, ou seja, linearizar-se. A expressão «a relação é linear» significa que a relação é linear ou linearizável relativamente aos parâmetros. (Murteira et al., 2015).

5.1 Coeficiente de correlação linear de Pearson:

"A correlação determina o grau em que duas variáveis estão relacionadas linearmente, seja por meio de causalidade direta, indireta ou por probabilidade estatística" (Barbetta, 2010).

Em primeiro lugar, para se proceder a uma análise de regressão, começar-se-á pelo cálculo da correlação, que pode ser feito através do coeficiente proposto por Pearson expresso da seguinte forma:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2(y_i - \bar{y})^2)}}$$

onde,

x_i e y_i = os valores das variáveis *X* e *Y*

$\bar{x}$ e $\bar{y}$= as médias dos valores x_i e y_i

O Coeficiente de Correlação Linear de Pearson (*r*) varia entre -1 e 1. Sendo que se este assumir o valor de 0 a correlação linear é nula, não existiria, portanto, relação linear entre duas variáveis quantitativas.

Se este coeficiente assumir um valor negativo quererá dizer que a correlação entre as variáveis é negativa, por outro lado, se assumir um valor positivo a correlação entre as variáveis é positiva. Esta será mais forte quanto mais próximo estiver dos valores 1 e -1 (Tabela 1).

Tabela 1 – Interpretação do coeficiente de correlação

Valor do coeficiente	Significado
$r = -1$	A correlação linear é negativa e perfeita.
$-1 < r \leq -0.8$	A correlação linear é negativa e forte.
$-0.8 < r \leq -0.5$	A correlação linear é negativa e moderada.
$-0.5 < r \leq -0.1$	A correlação linear é negativa e fraca.
$-0.1 < r \leq 0$	A correlação linear é negativa e ínfima.
$r = 0$	A correlação linear é nula, não havendo, portanto, relação linear entre as variáveis. Isto não significa que as variáveis sejam independentes, pois pode haver outro tipo de relação (não linear) entre as variáveis.
$0 \leq r < 0.1$	A correlação linear é positiva e ínfima.
$0.1 \leq r < 0.5$	A correlação linear é positiva e fraca.
$0.5 \leq r < 0.8$	A correlação linear é positiva e moderada.
$0.8 \leq r < 1$	A correlação linear é positiva e forte.
$r = +1$	A correlação linear é positiva e perfeita.

5.2 Coeficiente de determinação:

Uma forma de avaliar a adequabilidade do modelo aos dados consiste em dispor de um indicador que permita medir o 'grau de ajustamento' entre Y_i, as observações da variável dependente, e $\hat{Y}_i$, os respetivos valores ajustados *(*i= 1, 2, *…n).*

Para tal, será utilizado o Coeficiente de determinação,

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Os valores que este coeficiente pode tomar são: $0 \leq R^2 \leq 1$. O seu valor é zero quando nenhuma parcela da variação de Y é explicada pela regressão. Quanto mais próximo de 1 estiver este coeficiente melhor é o grau de ajustamento, ou seja, maior é a "proximidade" entre os y_i e os $\hat{y}_i$.

Neste estudo de Regressão Linear Múltipla, para analisar a qualidade de ajustamento de cada um dos modelos finais obtidos, será tido em conta o *adjusted* R^2, que o *software* estatístico apresentará, este R^2 irá considerar o número k de variáveis explicativas no modelo:

$$R_a^2 = R^2 - \frac{k(1-R^2)}{n-k-1}$$

5.3 Teste à significância global do modelo

Uma vez estimados os parâmetros da regressão, interessa saber se a relação observada entre as variáveis pode ser generalizada à população de onde as amostras foram recolhidas. Se isso acontecer, pode-se concluir que existe uma relação linear significativa entre as variáveis e a equação de regressão estimada, se se verificarem os pressupostos do modelo, pode ser utilizada na previsão dos valores da variável dependente.

No teste à significância global do modelo de regressão linear múltipla, as hipóteses são:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1: \exists\, \beta_j \neq 0 \end{cases}$$

A distribuição do teste a utilizar segue uma lei de F-Snedecor com *k* e *n-k-1* graus de liberdade:

$$F = \frac{MQR}{MQE} = \frac{\frac{SQR}{k}}{\frac{SQE}{n-k-1}} \sim F \qquad \text{ou} \qquad F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} = \frac{R^2(n-k-1)}{k(1-R^2)} = \sim F_{(k;n-k-1)}$$

5.4 Testes de significância individual

Quando se tem mais do que uma variável independente, e se pretende analisar a relação entre a variável dependente e cada uma das variáveis independentes, podem realizar-se testes de significância individuais aos parâmetros do modelo.

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

A estatística de teste, se a variância dos resíduos for desconhecida, é:

$$\frac{\hat{\beta}_j - \beta_j = 0}{\frac{s}{\sqrt{\sum x_{ji}^2}}} = \sim t_{(n-k-1)}$$

onde *k* é número de variáveis explicativas.

5.5 Pressupostos da Regressão Linear Múltipla:

Os pressupostos da regressão linear múltipla dizem respeito aos resíduos do modelo. A realização de uma análise residual é uma etapa essencial para investigação da adequabilidade do modelo desenvolvido através da regressão linear múltipla. Segundo Montgomery e Runger (2009), os resíduos têm um papel muito importante no julgamento do ajustamento do modelo.

Estes pressupostos são os seguintes:

- Os resíduos seguem uma Distribuição Normal
- Os resíduos têm valor médio nulo
- Homocedasticidade dos resíduos
- Independência dos resíduos
- Ausência de multicolinearidade

E podem ser apresentados como:

$\varepsilon_i \underset{\text{indep.}}{\sim} N(0, \sigma^2_\varepsilon)$ e ausência de multicolinearidade

Começo por recordar o que são os resíduos e como é que estes se obtêm. O resíduo representa a quantidade da variabilidade de Y que o modelo ajustado não consegue explicar.

Os resíduos podem ser calculados com a seguinte fórmula:

$$resíduo = Y_i - \hat{Y}_i$$

onde

Y_i = valor real (variável dependente)

$\hat{Y}_i$ = valor calculado pelo modelo (variável dependente ajustada)

5.5.1 Os resíduos seguem uma Distribuição Normal

As hipóteses a testar são:

$$\begin{cases} H_0: \varepsilon_i \sim N \\ H_1: \varepsilon_i \nsim N \end{cases}$$

É possível analisar este pressuposto graficamente, através, por exemplo, de um Q-Q Plot, onde os quantis da distribuição normal estarão representados por uma reta, e as observações representadas por pontos. Sendo que se estes se encontrarem muito afastados da reta suspeita-se que se rejeita o pressuposto da Normalidade.

De seguida, faz-se uma análise analítica através da realização de testes, como por exemplo o teste Shapiro-Wilk ou o teste Kolmogorov–Smirnov. De onde serão retirados os valores das estatísticas de teste e dos *p-values*.

Se este pressuposto falhar deverá proceder-se a uma transformação adequada dos dados, utilizando por exemplo a Transformação de Box-Cox.

5.5.2 Têm valor médio nulo

Pelo método dos mínimos quadrados é subentendido que a média dos resíduos é nula. De qualquer das formas, será feito o cálculo no *software* R para garantir que este pressuposto se encontra válido em todos os modelos apresentados.

5.5.3 Variância constante

Realiza-se o seguinte teste:

H_0: A variância dos ε_i é homogénea

H_1: A variância dos ε_i não é homogénea

Este pressuposto da homocedasticidade também é possível ser analisado tanto graficamente - através de um gráfico de dispersão dos resíduos (onde podemos observar se a dispersão dos resíduos é a mesma ao longo dos valores estimados, quando a variância graficamente é "constante", i.e., não existe uma aparente tendência de crescimento, os resíduos distribuem-se em torno de zero, ou, caso contrário, existe um problema de heterocedasticidade). Como analiticamente – através, por exemplo, do teste de Breusch-Pagan (utilizado quando é possível confirmar a Normalidade) ou de um Teste Score.

Caso o pressuposto de homocedasticidade não seja válido, pode-se dizer que os erros padrões dos estimadores, obtidos pelo Método dos Mínimos Quadrados, são incorretos e, portanto, a inferência estatística não é válida.

Tal como no pressuposto da Normalidade, se este pressuposto falhar deverá proceder-se a transformações aos dados.

5.5.4 Independência dos resíduos

Neste caso as hipóteses a testar são:

H_0: Não existe autocorrelação

H_1: Existe autocorrelação

Os resíduos devem ser independentes. A existência de autocorrelação é uma violação grave dos pressupostos do modelo linear, pois interfere diretamente na distribuição dos resíduos. Para a sua análise será utilizado o teste estatístico de Durbin-Watson.

5.5.5 Ausência de multicolinearidade

Por fim, numa RLM deve verificar-se se as variáveis independentes não se encontram fortemente correlacionadas, que não trazem a mesma informação ao modelo, ou seja, que não são redundantes. Para tal, verificam-se os valores de *VIF (Variance inflaction factor*) de cada variável explicativa do modelo. Estes valores de *VIF* devem ser inferiores a 10, caso contrário, significaria que existem variáveis fortemente correlacionadas no modelo e falha este pressuposto.

A falha deste pressuposto causa instabilidade na estimação dos coeficientes das variáveis do modelo, quer em termos de magnitude quer em termos de sinal. Assim sendo, das variáveis que são colineares, deve-se eliminar aquela que for menos significativa para o modelo e repetir a análise até que este pressuposto seja cumprido.

Nota: Ao longo deste trabalho o nível de significância base escolhido foi de 5%, ou seja, $\alpha = 0.05$, pelo que, em qualquer um dos pressupostos referidos, se *p-value* $> \alpha = 0.05$ não se rejeita H_0. E se *p-value* $\leq \alpha = 0.05$ rejeita-se H_0.

5.6 Existência de *outliers* ou observações influentes

Das várias opções utilizadas para a identificação de possíveis *outliers* no estudo serão analisados os valores de Leverage, um *outlier* será detetado se possuir um valor de Leverage > 0.5. Podemos designar *outlier* como um ponto que não se ajusta bem ao modelo, uma observação com valor discrepante de *Y*. Quando o valor ajustado $\hat{Y}_i$ é muito distante de Y_i, o resíduo ε_i é grande, e pode afetar o ajuste do modelo. Ao se

encontrar um *outlier* no estudo deve-se verificar também a possibilidade de ter ocorrido algum erro na recolha de dados.

A medida mais utilizada para detetar observações influentes, e que será utilizada neste estudo, é a Distância de Cook, que mede o quanto as estimativas dos coeficientes de regressão se alteram com a retirada de uma observação "i". Se a observação "i" tiver um valor de Distância de Cook superior a 1 indicar-nos-á que "i" é uma observação influente.

Uma observação influente, quando presente, altera substancialmente a qualidade do ajustamento do modelo. Daí a importância de incluir no estudo a análise da existência de observações influentes.

5.7 Métodos estatísticos de seleção de variáveis

Existem vários métodos que nos ajudam a decidir sobre a seleção de variáveis a incluir no modelo, de forma a chegar ao "melhor modelo". Estes métodos de seleção de variáveis apresentam a vantagem de indicar, com base num critério exato, quais as variáveis que apresentam relações mais fortes com a variável dependente e que por isso são melhores candidatas para o modelo definitivo.

Os três métodos que serão utilizados neste estudo, e aplicados através do pacote "caret" do *software* R, são os seguintes:

- *Forward:* Neste método o modelo inicial inclui apenas a constante. O *software* estatístico irá proceder à adição ao modelo de uma variável de cada vez conforme o teste *F* parcial, onde será escolhida a variável com maior coeficiente de correlação parcial, ou seja, o equivalente a ter maior valor de *F*, até ser apresentado o melhor modelo. Se o *F* parcial possuir um *p-value* menor que α, esta variável independente é adicionada ao modelo. Este procedimento continua até que uma determinada variável não possua um *F* com o *p-value* associado não seja menor que α, ou até que todas as variáveis independentes entrem no modelo.
- *Backward:* Neste método o modelo inicial inclui todas as variáveis independentes, a partir do qual são eliminadas, passo a passo, variáveis do modelo, também através do teste *F* parcial, até ser apresentado o

melhor modelo. A variável com maior *p-value* é comparada com o nível de significância fixado e se o valor do *p-value* for maior que o nível de significância, essa variável é removida do modelo. No próximo passo, um novo modelo com *k - 1* variáveis independentes é ajustado e o *p- value* do menor *F* parcial é comparado com o nível de significância fixado. Este procedimento continua até que não existam variáveis no modelo ou até que todas as variáveis presentes no modelo possuam um *p-value* associado ao *F* parcial inferior ao nível de significância.

- *Stepwise:* Este método será a conjugação dos dois métodos referidos acima. Inicia-se a seleção de variáveis apenas com uma variável independente, porém a significância de cada adição de uma nova variável ao modelo é testada como no método *Backward.* O teste *F* parcial continua a ser o teste utilizado para adicionar ou remover variáveis em cada passo.

Os diferentes métodos poderão conduzir a diferentes modelos sugeridos. Estes métodos serão um bom apoio para a identificação das variáveis que serão deixadas de fora pelos três métodos em simultâneo, contudo, a seleção do modelo final, na prática, passará também pelo conhecimento percecionado da temática sob estudo e pela discussão dos resultados com os colegas. Algo extremamente valioso no ajustamento de um modelo de aplicação prática.

5.8 Previsão

Uma análise através da regressão linear avalia se uma ou mais variáveis independentes explicam a variável dependente. Os modelos lineares são paramétricos, ou seja, possuem algumas suposições sobre os dados que analisam, como já foi referido. Quando essas suposições não são consideradas, os resultados da análise de regressão podem ser enganosos e o modelo pode não ter um bom desempenho.

"Até aqui, o modelo de regressão linear foi apresentado de acordo com o seguinte ponto de vista: dispondo de um certo número de observações sobre as variáveis, procura-se estimar uma relação linear capaz de explicar o comportamento do regressando (Y) em função de certos regressores (x_i). Outro ponto de vista é o de encontrar o modo mais eficaz de utilização do modelo com o **objetivo da previsão** de

observações adicionais do regressando a partir de certos valores assumidos pelos regressores. No entanto, deve sublinhar-se que só se deve passar à fase da previsão depois de se adotar um determinado modelo estimado, o que pressupõe que as estimações feitas foram submetidas a uma cuidada análise da especificação." (Murteira et al., 2015)

O «problema da previsão» procura dar resposta a dois tipos de questões:

a) Previsão pontual é uma estimativa de um único valor para um parâmetro populacional.

b) Previsão intervalar é um intervalo de valores usado para estimar um parâmetro populacional.

A previsão intervalar subdivide-se em previsão em média, quando se constrói um intervalo de confiança para a média da população, para um determinado conjunto de X_i, e previsão individual, quando se constrói um intervalo de confiança para uma observação isolada, para um determinado conjunto de X_i.

Neste caso o estudo em questão trata dados temporais, o que faz com que a previsão em média não tenha muito interesse, devido à própria natureza dos dados. Nas situações apresentadas existirá interesse em fazer previsão intervalar individual, isto é, prever um intervalo de confiança para apenas um particular valor da variável referido a outro período ou a outro contexto.

6. Modelação Estatística

Em seguida, passa-se para a fase de modelação estatística do estudo, onde serão apresentados quatro dos modelos experimentados ao longo do estágio. Para tal, foi utilizado o *software* estatístico R como ferramenta de apoio para a realização dos cálculos.

	Breve descrição do objetivo	Variável dependente
Primeiro Modelo	Análise da pobreza e desigualdades em Portugal.	Proporção da população residente em risco de pobreza ou exclusão social
Segundo Modelo	Análise da pobreza e desigualdades em Portugal.	Taxa de intensidade da pobreza
Terceiro Modelo	Estudo do impacto das respostas sociais na variação da variável dependente, relativamente ao nº de utentes.	Proporção da população residente em risco de pobreza ou exclusão social
Quarto Modelo	Estudo do impacto das respostas sociais na variação da variável dependente, relativamente ao nº de lugares.	Proporção da população residente em risco de pobreza ou exclusão social

Os dois primeiros modelos apresentados passarão pela modelação de duas variáveis diferenciadas, porém, ambas dentro da questão da análise da pobreza e desigualdades em Portugal. As variáveis independentes de ambos os modelos serão retiradas do Portal do Instituto Nacional de Estatística (INE) consideradas como indicadores de pobreza.

No Primeiro Modelo foi selecionada como variável dependente a "Proporção da população residente em risco de pobreza ou exclusão social", ou seja, a proporção de

indivíduos que vivem em Portugal em risco de pobreza ou em situação de privação material severa e/ou a viver em agregados com intensidade laboral *per capita* muito reduzida. Por outras palavras, pretende-se estudar a proporção de população que vive em Portugal que se encontra em situação de vulnerabilidade socioeconómica.

Em contrapartida, no Segundo Modelo alterou-se a variável dependente para: "Taxa de Intensidade da Pobreza", de forma a ser possível realizar uma análise da variação da mesma. Sendo que a intensidade da pobreza é um indicador que se destina a avaliar a medida em que o nível de vida da população abaixo do risco de pobreza está abaixo da linha de pobreza, ao ser analisada a intensidade da pobreza está a ser avaliada a intensidade desta vulnerabilidade socioeconómica.[13]

O Terceiro e Quarto Modelos apresentados incluirão, para além de dados do INE, elementos da Carta Social. Estes modelos passarão pela modelação da mesma variável dependente: "Proporção da população residente em risco de pobreza ou exclusão social". Nestes modelos pretende-se estudar o impacto da Rede de Serviços e Equipamentos Sociais (RSES), mencionada no início do relatório. Foram então selecionadas como variáveis independentes algumas das respostas sociais que integram da RSES para o grupo-alvo Família e Comunidade, disponíveis na Carta Social.

Como tal, no Terceiro Modelo serão analisados os dados que dizem respeito à frequência destas respostas sociais (o número de utentes em cada), isto é, a sua utilização. Enquanto que no Quarto Modelo, serão estudados os dados relativos à capacidade destas respostas sociais (o número de lugares que cada um faculta), ou seja, a sua cobertura de lugares. Pretende-se estudar de que forma as variáveis significativas que explicam a variação desta variável dependente serão as mesmas para ambos os modelos.

6.1 Primeiro Modelo:

Objetivo: modelar a "Proporção da população residente em risco de pobreza ou exclusão social" (PPRPES) através de nove variáveis representadas na Tabela 2.

[13] Para melhor compreensão do conceito ver ponto 6.2.

Tabela 2: Variáveis do modelo

Variáveis Independentes	Abreviaturas	Unidade
Carga mediana das despesas em habitação	CDH	%
Coeficiente de Gini	GINI	%
Desemprego registado	DR	Nº
Intensidade da privação material	IPM	Nº
Taxa de privação material	TPM	%
Taxa de sobrecarga das despesas em habitação	TDH	%
Índice de dependência total	IDT	Nº
Desigualdade na distribuição de rendimentos S80/S20	DDR80/20	Nº
Desigualdade na distribuição de rendimentos S90/S10	DDR90/10	Nº

→ **Variáveis do modelo:**

As variáveis foram selecionadas com a revisão da literatura e com o conhecimento empírico da equipa do Gabinete. Como já referido anteriormente, a variável dependente do estudo será: "Proporção da população residente em risco de pobreza ou exclusão social" = PPRPES, dentro do período de tempo: 2004-2020. Pois pretende-se estudar a relação desta variável com as seguintes variáveis independentes[14], retiradas do Portal do Instituto Nacional de Estatística (INE), e consideradas neste estudo como os indicadores de pobreza, ou seja, as variáveis que servem para explicar a variação nos valores da variável PPRPES:

→ Carga mediana das despesas em habitação: Indicador que traduz o rácio entre as despesas anuais associadas à habitação e o rendimento

[14] Definições retiradas do Portal do Instituto Nacional de Estatística. Disponíveis em: https://www.ine.pt

disponível do agregado, deduzindo as transferências sociais relativas à habitação em ambos os elementos da divisão.

(a) Despesas associadas a habitação: Despesas relacionadas com a renda, água, eletricidade, gás ou outros combustíveis, condomínio, saneamento, manutenção e pequenas reparações, bem como juros relativos ao crédito à habitação principal e seguros.

→ Coeficiente de Gini: Indicador de desigualdade na distribuição do rendimento que visa sintetizar num único valor a assimetria dessa distribuição. Assume valores entre 0 (quando todos os indivíduos têm igual rendimento) e 100 (quando todo o rendimento se concentra num único indivíduo).

→ Desemprego registado: Número de habitantes que fazem parte da população desempregada anualmente.

→ Intensidade da privação material: Média de itens de privação material em carência na população em situação de privação material.

→ Privação material: Condição do agregado doméstico privado no qual se verifica a carência forçada de pelo menos três dos seguintes nove itens, devido a dificuldades económicas: a) capacidade para assegurar o pagamento imediato de uma despesa inesperada e próxima do valor mensal da linha de pobreza (sem recorrer a empréstimo); b) capacidade para pagar uma semana de férias, por ano, fora de casa, suportando a despesa de alojamento e viagem para todos os membros do agregado; c) capacidade para pagar atempadamente rendas, prestações de crédito ou despesas correntes da residência principal, ou outras despesas não relacionadas com a residência principal; d) capacidade para ter uma refeição de carne ou de peixe (ou equivalente vegetariano), pelo menos de 2 em 2 dias; e) capacidade para manter a casa adequadamente aquecida; f) capacidade para ter máquina de lavar roupa; g) capacidade para ter televisão a cores; h) capacidade para ter telefone fixo ou telemóvel; i) capacidade para ter automóvel (ligeiro de passageiros ou misto).

→ Sobrecarga das despesas associadas à habitação: Condição dos agregados familiares cuja carga das despesas associadas à habitação é superior a 40%.

→ Índice de dependência total: Relação entre a população jovem e idosa e a população em idade ativa, definida habitualmente como o quociente entre o número de pessoas com idades compreendidas entre os 0 e os 14 anos conjuntamente com as pessoas com 65 ou mais anos e o número de pessoas com idades compreendidas entre os 15 e os 64 anos (expressa habitualmente por 100 (10^2) pessoas com 15-64 anos).

→ Desigualdade na distribuição de rendimentos S80/S20: Rácio S80/S20 é um indicador de desigualdade na distribuição do rendimento, definido como o rácio entre a proporção do rendimento total recebido pelos 20% da população com maiores rendimentos e a parte do rendimento auferido pelos 20% de menores rendimentos.

→ Desigualdade na distribuição de rendimentos S90/S10: Rácio S90/S10 é um indicador de desigualdade na distribuição do rendimento, definido como o rácio entre a proporção do rendimento total recebido pelos 10% da população com maiores rendimentos e a parte do rendimento auferido pelos 10% de menores rendimentos.

6.1.1 Estatística descritiva da variável dependente do modelo:

a) Medidas de tendência central:

As medidas de tendência central indicam, no geral, um valor central em torno do qual os dados estão distribuídos. As principais medidas de tendência central são as seguintes:

- Média aritmética, que indica o valor em torno do qual há um equilíbrio na distribuição dos dados e pode ser expressa da seguinte forma: $\bar{X} = \frac{\Sigma X}{n}$ ou seja, somar todos os valores observados da PPRPES e dividir pelo número de anos em estudo. "A média é calculada utilizando a magnitude dos valores, enquanto a mediana utiliza somente a ordenação dos valores." (Barbetta, 2010)

- Mediana, que é o valor central num conjunto de dados após ordenado. "A mediana

avalia o centro de um conjunto de valores, sob o critério de ser o valor que divide a distribuição ao meio, deixando os 50% menores valores de um lado e os 50% maiores valores do outro lado." (Barbetta, 2010)

- Moda, que é naturalmente o valor mais frequente no conjunto de dados.

  - Média = 24.436. A proporção da população residente em risco de pobreza ou exclusão social nos últimos 19 anos foi de 24.44%.
  - Mediana = 25. Sendo que os dados são temporais, no período de 2004 a 2020, em 50% dos anos a proporção da população residente em risco de pobreza ou exclusão social de foi no máximo 25%.

b) Medidas de dispersão:

Medidas de dispersão servem para quantificar a variabilidade dos valores num conjunto de dados. Existem diversas medidas de dispersão muito úteis e utilizadas, tais como:

- Amplitude total: É definida como a diferença entre a maior e a menor observação de um conjunto de dados.

- Desvio médio absoluto: É o cálculo da média dos desvios absolutos. Para este, deve ser primeiro calculada a média ($\bar{x}_{obs}$), posteriormente os desvios das observações em relação a média em módulo, e, por último, a média aritmética destes desvios, conforme a fórmula: $DMA = \frac{(\sum|X_i - \bar{X}|)}{n}$

- Variância: Esta quantifica a variabilidade dos dados em torno da média, ou seja, mede o quão longe do valor esperado os dados se encontram. O seu cálculo é feito através da seguinte fórmula: $S^2 = \frac{\sum(X - \bar{X})^2}{n-1}$

- Desvio padrão: Como se pode perceber pela fórmula apresentada abaixo, o cálculo da variância eleva ao quadrado a soma da diferença dos desvios. Isto faz com que a variância não tenha a interpretação na mesma escala em que os dados foram medidos. Para contornar isso faz-se a raiz quadrada da variância para que este novo valor esteja na mesma escala e se possa analisar a variabilidade dos dados na mesma escala em que

eles foram medidos. “A média e o desvio padrão são as medidas mais usadas para avaliar a posição central e a dispersão de um conjunto de valores. Contudo, essas medidas são fortemente influenciadas por valores discrepantes.” (Barbetta, 2010)

$$S = \frac{\sqrt{\sum(X - X)^2}}{n - 1}$$

- Coeficiente de variação: É a razão entre o desvio padrão e a média. Esta é uma medida relativa que avalia o percentual de variabilidade em relação a média observada. Uma das grandes vantagens desta medida é a possibilidade de comparar a variabilidade de conjuntos medidos em diferentes escalas e ainda observar o comportamento da variabilidade de uma variável quantitativa entre as categorias de uma variável qualitativa e calcula-se da seguinte forma:

$$CV = \frac{S}{X} * 100\%$$

Na Figura 2 pode-se observar o cálculo das medidas de dispersão apresentadas em relação à Proporção da população residente em risco de pobreza ou exclusão social = PPRPES, variável dependente deste modelo.

amplitude	desvio_medio_abs	variancia	desvio_padrão	coef_var.
8.1	37.11579	6.07	2.46	10.07

Figura 2 - Medidas de dispersão da variável PPRPES

- Amplitude total = 8.1. A diferença entre a maior Proporção da população residente em risco de pobreza ou exclusão social (PPRPES) e a menor, nos últimos 19 anos, é de 8.1%
- Desvio médio absoluto = 37.12. A distância a que os dados tendem a estar do valor central – média – é de 37.12. Isto é, a média dos desvios absolutos entre os valores de PPRPES e a média é 37.12.
- Variância = 6.07. Tanto a variância quanto o desvio padrão são medidas que fornecem informações complementares à informação da média aritmética. Estas medidas avaliam a dispersão do conjunto de valores em análise, porém, fará mais sentido interpretar o desvio-padrão, pois tem a unidade de medida da variável, sendo que a variância é a unidade de medida ao quadrado.

- Desvio padrão = 2.46. O desvio típico em relação à média e de 2.46%. Sobre a variabilidade dos dados em relação à média, um desvio padrão mais alto indica uma maior dispersão nos dados, neste caso, 2.46% é um valor que indica que os dados não se encontram muitos dispersos em relação à média.
- Coeficiente de variação = 10.07%. Para valores inferiores a 50% do coeficiente de variação (ou 0,5 do coeficiente de dispersão) a média será tanto mais representativa quanto menor o valor deste coeficiente, e, assim sendo, quanto menor for este valor mais homogéneo é o conjunto de dados. Consequentemente, valores superiores a 50% do coeficiente de variação indicam uma pequena representatividade da média. Como este valor é de 10.07%, diferente de zero, significa que há variabilidade entre os dados, tal como, também não sendo superior a 50%, significa que a média é representativa.

6.1.2 Análise de Correlação

De seguida, apresentam-se os Coeficientes de Correlação de Pearson da variável dependente com cada uma das variáveis de estudo independentes, obtidos através do *software* R:

Tabela 3: Coeficientes de Correlação de Pearson

Variáveis independentes	**Coeficiente de Correlação de Pearson**	**Interpretação**	***p-value***
Carga mediana das despesas em habitação	0.352	Correlação fraca positiva	0.014
Coeficiente de Gini	0.662	Correlação moderada positiva	0.003
Desemprego registado	0.697	Correlação moderada positiva	0.001
Intensidade da privação material	0.688	Correlação moderada	0.002

		positiva	
Taxa de privação material	0.926	Correlação muito forte positiva	≈ 0
Taxa de sobrecarga das despesas em habitação	0.476	Correlação fraca positiva	0.039
Índice de dependência total	-0.658	Correlação moderada negativa	0.004
Desigualdade na distribuição de rendimentos s80/s20	0.705	Correlação forte positiva	0.001
Desigualdade na distribuição de rendimentos s90/s10	0.726	Correlação forte positiva	0.001

Sendo que os *p-values* dos testes realizados à significância da correlação são todos valores pequenos (< 0.05), conclui-se que as correlações são todas significativas.

6.1.3 Modelação:

Neste Primeiro Modelo a equação a ajustar é a seguinte:

$PPRPES_i = \beta_0 + \beta_1 CDH + \beta_2 GINI + \beta_3 DR + \beta_4 IPM + \beta_5 TPM + \beta_6 TDH + \beta_7 IDT + \beta_8 DDR80/20 + \beta_9 DDR90/10 + \varepsilon_i$

Parâmetro	Estimate	T_{obs}	*p-value*
β_0	-2.865e+01	-0.552	0.598
β_1	-1.601e-01	-0.293	0.778
β_2	-2.495e-01	-0.204	0.844
β_3	-2.562e-06	-0.601	0.567
β_4	1.001e+01	1.829	0.110
β_5	5.744e-01	3.898	0.006
β_6	5.006e-02	0.109	0.916
β_7	1.838e-01	0.374	0.720
β_8	-4.041e-01	-0.087	0.933
β_9	9.512e-01	0.514	0.623

Verificação dos Pressupostos ao modelo inicial:

Os resíduos do modelo inicial são independentes, seguem uma distribuição normal, com média nula, variância constante, porém, existem problemas de multicolinariedade. Existem seis variáveis com *VIF* (*Variance Inflation Factor*) superior a 10, o que significa que várias das nossas variáveis independentes se encontram fortemente correlacionadas, trazendo possivelmente informação redundante ao modelo. Assim sendo, de forma a validar este pressuposto fulcral, começou-se por remover uma das variáveis com *VIF*>10, aquela que seria menos significativa com base nos *p-values* dos testes t.

CDH	GINI	DR	IPM	TPM	TDH	IDT	`DDR80/20`	`DDR90/10`
15.141544	153.413213	6.627309	4.415017	6.527705	17.963553	32.794049	166.358157	115.596487

Figura 3 - Valores VIF do Modelo inicial

A primeira variável a ser eliminada foi: `DDR80/20`, cujo *p-value*= 0.933. Obteve-se um novo modelo – "Modeloa". Contudo, continuam a existir problemas de multicolinearidade. Como é possível observar na Figura 4.

CDH	GINI	DR	IPM	TPM	TDH	IDT	`DDR90/10`
8.590188	119.235119	6.543898	3.776820	6.409999	10.002144	32.356402	56.109645

Figura 4 - Valores VIF do Modeloa

De seguida, foram feitas várias experiências. Uma destas passou por eliminar a variável TDH, que mostrava ter o maior *p-value* (0.818), o que fazia com que permanecessem três variáveis no modelo com *VIF*>10. Passando, pela mesma lógica, para a eliminação da variável IDT, o que fazia com que duas variáveis ainda tivessem um *VIF*>10: GINI e `DDR90/10`. Entre estas, permaneceu a variável `DDR90/10` (*p- value*= 0.026) e foi eliminada a variável GINI (*p-value*= 0.104), obtendo-se, por fim, um modelo onde todos os pressupostos se encontram válidos.

O estudo continuará então com o "Modelod":

$PPRPES = \beta_0 + \beta_1 CDH + \beta_2 DR + \beta_3 IPM + \beta_4 TPM + \beta_5 DDR90/10 + \varepsilon_i$

Utilização dos métodos estatísticos de seleção de variáveis:

De seguida, utilizaram-se os métodos estatísticos de seleção de variáveis: *Backward, Forward* e *Stepwise*, de forma a obter o seguinte modelo:

$PPRPES_i = \beta_0 + \beta_1 IPM + \beta_2 TPM + \beta_3 DDR90/10 + \varepsilon_i$

Modelo ajustado:

$\widehat{PPRPES}_i = -7.58 + 5.10\, IPM_i + 0.48\, TPM_i + 0.41\, DDR90/10_i$

Testar a significância individual:

Para estudar a significância individual utilizamos os testes t:

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

$i = 1, 2, 3, 4$

	T_{obs}	*p-value*
β_1	1.72	$0.110 > \alpha$
β_2	7.43	$\approx 0 < \alpha$
β_3	2.09	$0.057 > \alpha$

Decisão: Rejeitar H_0 para β_1, β_3 e β_4.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, existe evidência estatística de que as variáveis TPM e DDR90/10 são significativas individualmente. Porém, a variável IPM tem um *p-value* > $\alpha = 0.05$, o que significa que esta não é significativa individualmente. Assim sendo, esta variável será retirada do modelo.

Modelo final a ajustar:

$PPRPES_i = \beta_0 + \beta_1 TPM + \beta_2 DDR90/10 + \varepsilon_i$

Modelo final ajustado:

$\widehat{PPRPES}_i = 8.73 + 0.52\, TPM + 0.54\, DDR90/10$

```
> summary(modelodfinal)

Call:
lm(formula = PPRPES ~ TPM + `DDR90/10`)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2204 -0.4278 -0.1563  0.5490  1.2298

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.73442    1.64350   5.315 0.000109 ***
TPM          0.51695    0.06448   8.017 1.34e-06 ***
`DDR90/10`   0.54155    0.19227   2.817 0.013723 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7053 on 14 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.9093,    Adjusted R-squared:  0.8963
F-statistic: 70.16 on 2 and 14 DF,  p-value: 5.058e-08
```

Figura 7 - Modelod final

Qualidade do ajustamento:

$R_a^2 = 0.89$

A qualidade do ajustamento é muito boa. 89% da variabilidade total da percentagem da população residente em risco de pobreza ou exclusão social (PPRPES) é explicada pela equação ajustada.

Significância global do modelo:

Teste F da ANOVA:

$H_0: \beta_1 = \beta_2 = 0$

$H_1:] \beta_i \neq 0$

$i = 1, 2$

$F_{obs} = 70.16$

p-value aproximadamente $0 < \alpha = 0.05$ Decisão: Rejeitar H_0

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, existe evidência estatística de que o modelo é globalmente significativo.

Interpretação dos coeficientes ajustados das variáveis significativas:

$\hat{\beta}_0 = 8.734$: ordenada na origem, só faria sentido interpretar se as restantes variáveis assumissem o valor zero.

$\hat{\beta}_1 = 0.517$: Se a Taxa de privação material (TPM) aumentar um ponto percentual, estima-se que a % da população residente em risco de pobreza ou exclusão social aumente em média 0.517%, *ceteris paribus*.

$\hat{\beta}_2 = 0.542$: Se (DDR90/10) o Rácio S90/S10 (indicador de desigualdade na distribuição do rendimento), aumentar um valor, estima-se que a % da população residente em risco de pobreza ou exclusão social aumente em média 0.542%, *ceteris paribus*.

Resultados do Primeiro Modelo:

É possível concluir que a questão da privação material está diretamente correlacionada com a variação da proporção de população residente em Portugal em risco de pobreza ou exclusão social.

Pode-se entender como definição de intensidade da privação material, a média de itens de privação material em carência na população em situação de privação material. [15]

Apesar de muitos autores afirmarem que a pobreza não deve ser definida apenas como carência material, mas numa perspetiva mais abrangente no âmbito de um desenvolvimento humano pleno e que afirma que é maior pobreza "não ser socialmente" do que "não ter" (Mendes, et al., 2005), grande parte dos estudos sobre pobreza enfatizam a referida carência material.

Assim, pode entender-se o facto das duas variáveis independentes que tratam a dimensão da privação material se manterem ambas nos modelos ao longo do estudo, e da variável Taxa de privação material (TPM) se mostrar como a variável mais significativa para o modelo final pelo quanto afeta a variação da variável dependente - proporção de população residente em Portugal em risco de pobreza ou exclusão social. Assumindo-se que a privação material resulta explicitamente de uma incapacidade com origem financeira, não estando associada a uma escolha livre dos indivíduos, esta situação conduz a estados de carência por parte das famílias ou indivíduos definidos como pobres. A ideia subjacente à definição de pobreza é, portanto, a de possuir menos do que o mínimo imprescindível à satisfação das necessidades básicas (Mendes, et al., 2005).

Pode-se concluir também que, dos vários rácios indicadores de desigualdade, o Rácio S90/S10 é aquele que se mostra mais significativo para explicar a variação da proporção de pessoas em Portugal em risco de pobreza ou exclusão social. O que faz sentido sendo que Portugal é um dos países europeus que apresenta um nível de maior concentração do rendimento disponível auferido no grupo dos 10% do topo.[16] Em

---

[15] E, tal como afirma Nuno Alves no seu artigo "Um indicador de pobreza multidimensional para Portugal", publicado em 2022, no domínio de privação material incluem-se os seguintes indicadores: Incapacidade de pagamento de uma despesa inesperada; Incapacidade de pagar uma semana de férias fora de casa; Incapacidade de cumprimento de compromissos financeiros; Incapacidade de ter uma alimentação adequada; Indisponibilidade de viatura própria; Incapacidade de substituir vestuário ou calçado; E a privação de um computador por dificuldades económicas.

[16] Leitura "Rácio S90/S10". Disponível em: https://observatoriodasdesigualdade.wordpress.com

consequência, para analisar a evolução desta variável dependente deve ter-se em atenção a evolução do valor do rácio entre a proporção do rendimento total recebido pelos 10% da população com maiores rendimentos e a parte do rendimento auferido pelos 10% da população de menores rendimentos.

Verificação dos pressupostos do modelo final:

- Normalidade:

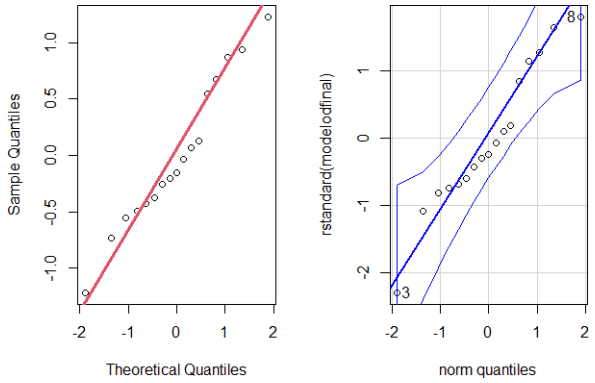

Figura 8 - Normalidade dos Resíduos

Para se analisar a normalidade dos resíduos graficamente, construíram-se dois gráficos de Probabilidade Normal dos Resíduos (Figura 8), onde se pode observar que estes seguem aproximadamente uma linha reta e que os pontos se encontram praticamente todos dentro da banda azul, ou seja, da banda de confiança de 95%, à exceção de dois valores que se encontram mais afastados, pelo que não parecem existir problemas com a validação da normalidade.

Assim sendo, para confirmar a interpretação acima, este pressuposto foi analisado analiticamente, através do teste de Kolmogorov-Smirnow e do teste de Shapiro-Wilk.

```
> lillie.test(resid(modelodfinal))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(modelodfinal)
D = 0.12805, p-value = 0.6475

> shapiro.test(resid(modelodfinal))

        Shapiro-Wilk normality test

data:  resid(modelodfinal)
W = 0.96527, p-value = 0.7315
```

Tanto no teste de Kolmogorov-Smirnow como no teste de Shapiro-Wilk, vemos que os *p-values* são superiores a $\alpha = 0.05$, o que nos leva a não rejeitar a hipótese nula.

Conclui-se que, com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que os resíduos não seguem uma Distribuição Normal.

- Média nula:

Pelo *software* R confirma-se que a média dos resíduos é aproximadamente zero.

- Homocedasticidade:

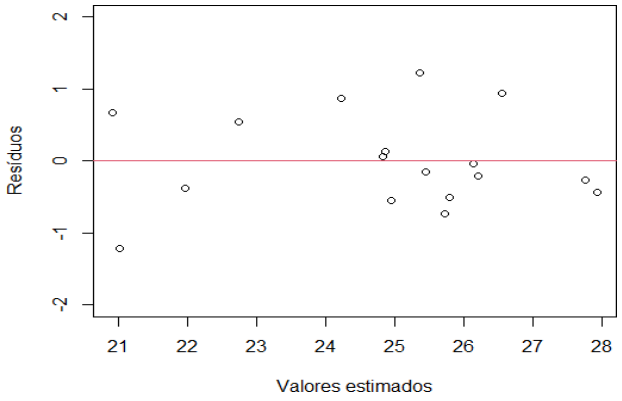

Figura 9 - Resíduos vs Valores estimados

Ao se analisar graficamente através da Figura 9 afirma-se que a variância é constante pois os valores encontram-se dispersos e não seguem propriamente um padrão de aumento/diminuição da dispersão.

Para uma análise mais completa realizou-se o Teste de Breusch-Pagan:

```
> bptest(modelodfinal)

        studentized Breusch-Pagan test

data:  modelodfinal
BP = 4.0288, df = 2, p-value = 0.1334
```

$BP_{obs} = 4.028$

p-value= $0.133 > \alpha = 0.05$

Decisão: Não se rejeita H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, conclui-se que não existe evidência estatística de que a variância dos resíduos não é homogénea, o pressuposto está validado.

- Independência dos resíduos:

Para verificar o pressuposto da independência dos resíduos realizou-se o Teste de Durbin-Watson:

```
> dwtest(modelodfinal, alternative = "two.sided")

        Durbin-Watson test

data:  modelodfinal
DW = 1.4798, p-value = 0.111
alternative hypothesis: true autocorrelation is not 0
```

DW_{obs}= 1.480

p-value= 0.111 > α = 0.05

Decisão: Não rejeitar H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que existe autocorrelação.

- Ausência de Multicolinearidade:

```
      TPM `DDR90/10`
 1.443019   1.443019
```

Ambos os VIF < 10

Conclui-se que todos os pressupostos são válidos, é possível utilizar o modelo para fazer previsão.

Pesquisa de observações influentes:

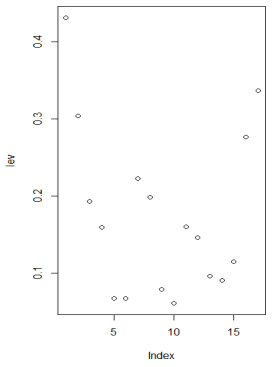

Figura 10 - Valores de Leverage

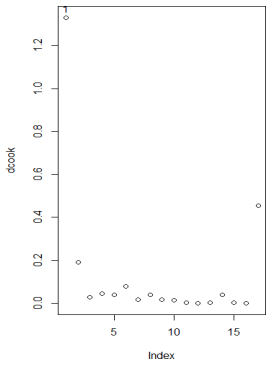

Figura 11 - Distância de Cook

É possível observar no gráfico da Figura 10 que a observação 1 se encontra com um valor de Leverage ligeiramente mais elevado que as restantes observações, porém, esta não será considerada um *outlier*, visto que este valor não é superior a 0.5, assim sendo, não afetará o ajuste do modelo.

Após revisão da base de dados conclui-se que não parece que haja erros de registo, sendo que todos os pontos são observações reais, não será retirada nenhuma observação. A distância de Cook desta observação – representada no gráfico da Figura 11– apesar de nitidamente mais elevada que as restantes observações, é de 1.32, ou seja, não é muito superior a 1.

Previsão:

Com o *software* estatístico R foram realizados cálculos de previsão através do modelo final obtido.

Nesta situação existe interesse em fazer previsão intervalar, isto é, prever possíveis valores para a variável dependente alterando os valores das variáveis independentes do modelo final.

Começa-se por observar o intervalo de variação das variáveis independentes, pois é aconselhada a interpolação, a previsão de valores diferentes da amostra, mas minimamente dentro dos valores reais, de forma a que o novo cenário hipotético não seja um cenário irrealista.

Como é possível observar na Tabela 4, no ano de 2020, o valor real da variável dependente está contido no intervalo de predição calculado, pelo que se pode concluir que o modelo estará a funcionar corretamente.

Para o ano de 2021 o valor da variável independente "TPM" não se encontrava ainda disponível na base de dados, pelo que este foi estipulado/escolhido conforme as leituras feitas e discussão do tema, criando um cenário onde esta taxa de privação material ("TPM") subiu, de forma a ir ao encontro da subida do valor real de "PPRPES" nesse ano.

Tabela 4: Previsão da variável dependente PPRPES

Ano	Valor pontual	Intervalo de Confiança (95%)	Valor real		
	estimado para "PPRPES"	individual de predição de nova observação	de "PPRPES"	"TPM"	"DDR90/10"
2020	21.02%	(19.21 ; 22.83)	19.8%	13.5*	9.8*
2021	20.83%	(19.12 ; 22.55)	22.4%	14.5	8.5*

*= Estes dados são valores reais retirados do Portal do INE.

Na Tabela 5 pretende-se apresentar uma experiência realizada onde a "TPM" se mantém a 13.5% enquanto o Rácio S90/S10 ("DDR90/10") continua em decréscimo, como mostram os seus valores reais, (sendo este um indicador de desigualdade na distribuição do rendimento, quanto menor o seu valor menor esta desigualdade). O que levaria a uma muito boa previsão do valor da variável dependente do estudo, como é possível observar na Tabela 5, o valor pontual estimado para 2022 está muito próximo do valor real que se obteve posteriormente ao estudo.

Depois, no seguimento deste cenário otimista, experimentou-se para o ano de 2023, a variável "DDR90/10" continuar em decréscimo e a variável "TPM" apresentar também uma descida. Desta forma a variável dependente iria continuar a decrescer significativamente.

Por fim, apresenta-se o cenário hipotético para 2024, onde a variável "DDR90/10" se mantém no valor hipotético: 7.5 e, desta vez, é a variável "TPM" que apresenta um decréscimo para 1 ponto percentual abaixo do valor mínimo da sua amplitude. O que resultaria numa diminuição da predição de "PPRPES", porém, não tão acentuada como no cenário para o ano de 2023.

Tabela 5: Cenário otimista

Ano	Valor pontual estimado para "PPRPES"	Intervalo de Confiança (95%) individual de predição de nova observação	Valor real de "PPRPES"	"TPM"	"DDR90/10"
2022	20.05%	(18.26 ; 21.83)	20.1%	13.5	8
2023	19.52%	(17.67 ; 21.37)		13	7.5
2024	19.26%	(17.39 ; 21.13)		12.5	7.5

6.2 Segundo Modelo:

O intuito do Segundo Modelo apresentado passou pela experiência de alteração da própria variável dependente. Esta manteve-se ao longo do estudo como a proporção de população residente em Portugal em risco de pobreza ou exclusão social, porém, neste modelo decidiu-se utilizar a variável "Taxa de intensidade da pobreza" (TIP) como variável resposta e realizar uma análise da variação da mesma.

- Definição de Intensidade da pobreza: Indicador que se destina a avaliar a medida em que o nível de vida da população abaixo do risco de pobreza está abaixo da linha de pobreza e que se calcula da seguinte forma:

$$\frac{Linha\ de\ pobreza - Rendimento\ médio\ da\ população\ abaixo\ da\ linha\ de\ pobreza}{Linha\ de\ pobreza}$$

- Definição de linha de pobreza: Limiar do rendimento abaixo do qual se considera que uma família se encontra em risco de pobreza. Este valor foi convencionado pela Comissão Europeia como sendo o correspondente a 60% [17] da mediana do rendimento por adulto equivalente de cada país. [18]

Objetivo: modelar a variável Taxa de intensidade da pobreza (TIP) através de 12 variáveis representadas na Tabela 2 e Tabela 6.

Neste modelo as variáveis independentes serão as variáveis utilizadas no Primeiro Modelo, que se encontram na Tabela 2, com acrescento das seguintes variáveis que se podem observar na Tabela 6:

Tabela 6: Variáveis do modelo

Variáveis Independentes	Abreviaturas	Unidade
População residente em risco de pobreza ou exclusão social	PPRPES	%
Taxa de risco de pobreza antes das transferências sociais	TRPA	%

[17] Ainda que possam ser disponibilizadas outras percentagens.

[18] Definições retiradas do Portal do Instituto Nacional de Estatística. Disponíveis em: https://www.ine.pt

Taxa de risco de pobreza depois das transferências sociais	TRPD	%

→ **Variáveis do modelo:**

Como já referido anteriormente, a variável depende do estudo será: "Taxa de intensidade da pobreza" = TIP, dentro do período de tempo escolhido para este estudo: 2003-2021. Pois pretende-se estudar a relação desta variável com as diversas variáveis independentes, retiradas do Portal do Instituto Nacional de Estatística (INE), incluindo como variável explicativa a "Proporção da população residente em risco de pobreza ou exclusão social" = PPRPES, de forma a ser possível analisar que variáveis servem para explicar a variação nos valores desta Taxa de intensidade da pobreza. Assim sendo, como referido na Tabela 6, foram adicionadas ao modelo outras duas variáveis distintas: a Taxa de risco de pobreza antes das transferências sociais; e a Taxa de risco de pobreza depois das transferências sociais, que dizem respeito à evolução do risco de pobreza com a existência de "transferências sociais", que passarão a ser definidas:

→ Transferências sociais em sentido lato: "Transferências sociais que correspondem às pensões provenientes de planos individuais, privados ou públicos (prestações de velhice e sobrevivência) e outras relativas a família, educação, habitação, doença/invalidez, desemprego e combate à exclusão social." [19]

6.2.1 Estatística descritiva da variável dependente:

a) Medidas de tendência central:

- Média = 25.04. Em média a Taxa de intensidade de pobreza de 2003 a 2022 foi de 25.04%.
- Mediana = 24.40. Sendo que os dados são temporais, no período de 2004 a 2022, em 50% dos anos a Taxa de Intensidade de Pobreza foi no máximo 24.40%.

19 Definições retiradas do Portal do Instituto Nacional de Estatística. Disponíveis em: https://www.ine.pt

b) Medidas de dispersão:

Na Figura 12 pode-se observar o cálculo das medidas de dispersão apresentadas em relação à Taxa de intensidade da pobreza= TIP, variável dependente deste modelo. Como TIP é um variável quantitativa, todas as medidas de dispersão apresentadas podem ser calculadas.

amplitude	desvio_medio_abs	variancia	desvio_padrão	coef_var.
8.6	36.41053	5.43	2.33	9.3

Figura 12 - Medidas de dispersão da variável TIP

- Amplitude total = 8.6. A diferença entre a maior e a menor Taxa de intensidade de pobreza nos últimos 19 anos é de 8.6%
- Desvio médio absoluto = 36.41. A média dos desvios absolutos entre os valores de TIP e a média é 36.41.
- Variância = 5.43.
- Desvio padrão = 2.33. No caso da variância e desvio-padrão fará mais sentido interpretar o desvio-padrão, pois tem a unidade de medida da variável, sendo que a variância é a unidade de medida ao quadrado. O desvio típico em relação á média é de 2.33%.
- Coeficiente de variação = 9.3. Como este valor é de 9.3%, diferente de zero, significa que há variabilidade entre os dados, tal como, também não sendo superior a 50%, significa que a média é representativa.

6.2.2 Análise de Correlação:

De seguida, apresenta-se um gráfico de correlação obtido através do *software* R, como se observa na Figura 13, que permite identificar os Coeficientes de Correlação de Pearson entre cada uma das variáveis do estudo e, consequentemente, se estas se correlacionam de forma forte ou fraca, positiva ou negativa. Sendo que o que nos interessará principalmente será a correlação entre a variável dependente e cada uma das variáveis independentes do estudo.

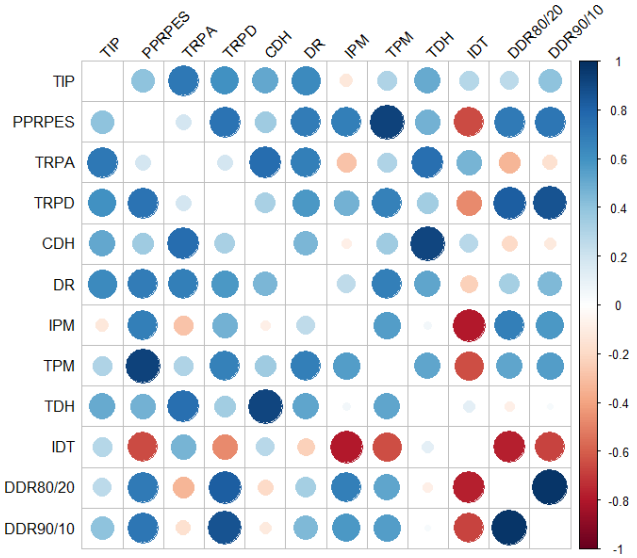

Figura 13 - Gráfico de Correlação

É possível concluir através da primeira coluna do gráfico da Figura 13 que, das diversas variáveis independentes apresentadas, a variável "IPM" é a única que mostra ter uma correlação negativa com a variável dependente Taxa de Intensidade da Pobreza, sendo esta uma correlação muito fraca.

Ao mesmo tempo, é possível interpretar também que as variáveis "TRPA", "TRPD" "DR" são as variáveis que estão correlacionadas de forma mais forte, e positiva, com a variável dependente deste modelo.

6.2.3 Modelação:

No Segundo Modelo a equação a ajustar é a seguinte:

$$TIP_i = \beta_0 + \beta_1\, PPRPES + \beta_2\, TRPA + \beta_3\, TRPD + \beta_4\, CDH + \beta_5\, DR + \beta_6\, IPM + \beta_7\, TPM + \beta_8\, TDH + \beta_9\, IDT + \beta_{10}\, DDR80/20 + \beta_{11}\, DDR90/10 + \varepsilon_i$$

Verificação dos Pressupostos ao modelo inicial:

Os resíduos são independentes, seguem uma distribuição normal, com média nula, variância constante, porém, falha totalmente o pressuposto da ausência de multicolinariedade. Todas as onze variáveis têm valores de *VIF* (*Variance Inflaction Factor*) superiores a 10. O que causa uma enorme instabilidade na estimação dos coeficientes de todas as variáveis do modelo, tanto na sua magnitude como no seu sinal. Desta forma não seria possível a utilização do modelo para quaisquer conclusões.

Começar-se-á por se retirar a variável que se mostra menos significativa,

conforme os valores de *p-value*, que, inicialmente, será a variável "Taxa de risco de pobreza depois das transferências sociais" = TRPD, como se pode reparar na Figura 14. Este processo de eliminação da variável menos significativa irá repetir-se até à chegada de um modelo onde o pressuposto da multicolinearidade esteja válido.

```
> summary(modelosegundo)

Call:
lm(formula = TIP ~ PPRPES + TRPA + TRPD + CDH + DR + IPM + TPM +
    TDH + IDT + `DDR80/20` + `DDR90/10`)

Residuals:
       3        4        5        6        7        8        9       10       11       12
 0.57769  0.91915 -1.77039  0.08996  0.56041 -0.99005  0.11590  0.71381 -0.20017 -0.49747
      13       14       15       16       17       18       19
-0.48153  0.86742  0.24620  0.12618  0.24052 -0.54076  0.02313

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.390e+01  6.814e+01  -0.791    0.465
PPRPES       7.168e-02  7.925e-01   0.090    0.931
TRPA        -3.472e-01  1.117e+00  -0.311    0.769
TRPD         1.245e-01  1.876e+00   0.066    0.950
CDH          3.691e-01  1.087e+00   0.340    0.748
DR           1.257e-05  1.504e-05   0.835    0.442
IPM         -6.045e+00  1.407e+01  -0.430    0.685
TPM          1.956e-01  6.530e-01   0.299    0.777
TDH          1.415e-01  9.928e-01   0.143    0.892
IDT          1.385e+00  8.615e-01   1.607    0.169
`DDR80/20`   5.529e+00  7.035e+00   0.786    0.468
`DDR90/10`  -8.234e-01  3.202e+00  -0.257    0.807

Residual standard error: 1.252 on 5 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.9085,    Adjusted R-squared:  0.7072
F-statistic: 4.514 on 11 and 5 DF,  p-value: 0.05431
```

Figura 14 - Modelo 2 - Modelo inicial

Após este passo chegou-se ao modelo:

$TIP_i = \beta_0 + \beta_1 CDH + \beta_2 DR + \beta_3 IPM + \beta_4 TPM + \beta_5 IDT + \beta_6 DDR80/20 + \varepsilon_i$

Neste modelo todos os pressupostos da regressão linear múltipla encontram-se válidos. Porém, no modelo representado acima existem muitas variáveis que não são significativas individualmente.

Testar a significância individual:

Para estudar a significância individual utilizamos os testes t:

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

i = 1, 2, 3, 4, 5, 6

	T_{obs}	*p-value*
β_1	1.267	0.233857 > α
β_2	2.358	0.040108 < α
β_3	-0.653	0.528250 > α
β_4	1.253	0.238590 > α

β_5	4.491	0.001159 < α
β_6	5.412	0.000297 < α

Decisão: Rejeitar H_0 para β_2, β_5 e β_6.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, existe evidência estatística de que as variáveis DR, IDT, DDR80/20 e CDH são significativas individualmente. Porém, as variáveis IMP e TPM tem um *p-value* > α = 0.05, o que significa que estas não são estatisticamente significativas individualmente. Será construído um novo modelo.

$TIP_i = \beta_0 + \beta_1 CDH + \beta_2 DR + \beta_3 IDT + \beta_4 DDR80/20$

```
> summary(modelosegundof)

Call:
lm(formula = TIP ~ CDH + DR + IDT + `DDR80/20`)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6944 -0.5523  0.1684  0.3337  1.1936

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.273e+01  1.155e+01  -5.433 0.000152 ***
CDH          5.142e-01  2.098e-01   2.451 0.030521 *
DR           9.027e-06  2.514e-06   3.591 0.003708 **
IDT          1.039e+00  1.582e-01   6.563 2.68e-05 ***
`DDR80/20`   3.948e+00  7.001e-01   5.638 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9167 on 12 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.8823,    Adjusted R-squared:  0.843
F-statistic: 22.48 on 4 and 12 DF,  p-value: 1.676e-05
```

Figura 15 - Modelo 2 - Modelo final

Como é possível confirmar pela Figura 15, no modelo final todas as variáveis se mostram significativas. Após a aplicação dos métodos estatísticos de seleção de variáveis: *Backward, Forward* e *Stepwise*, confirma-se o modelo final apresentado.

Modelo final ajustado:

$\hat{TIP}i = - 62.73 + 0.5142\, CDH + 0.000009027\, DR + 1.039\, IDT + 3.948\, DDR80/20$

Qualidade do ajustamento:

$R_a^2 = 0.843$

A qualidade do ajustamento é muito boa. 84% da variabilidade total da Taxa de intensidade da pobreza (TIP) é explicada pela equação ajustada.

Significância global do modelo:

Teste F da ANOVA:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1:] $\beta_i \neq 0$

i = 1, 2, 3, 4

$F_{obs} = 22.48$

p-value aproximadamente $0 < \alpha = 0.05$ Decisão: Rejeitar H_0

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, existe evidência estatística de que o modelo é globalmente significativo.

Interpretação dos coeficientes ajustados das variáveis significativas:

$\hat{\beta}_0 = -62.73$: ordenada na origem, só faria sentido interpretar se as restantes variáveis assumissem o valor zero.

$\hat{\beta}_1 = 0.5142$: Se (CDH) a carga mediana das despesas associadas à habitação aumentar um ponto percentual estima-se que a Taxa de intensidade da pobreza aumente em média 0.51%, *ceteris paribus*.

$\hat{\beta}_2 = 0.000009027$: Se (DR) o desemprego registado aumentar uma unidade, estima-se que a Taxa de intensidade da pobreza aumente em média 0.000009027%, *ceteris paribus.*

$\hat{\beta}_3 = 1.039$: Se (IDT) o Índice de Dependência Total aumentar um valor, estima-se que a Taxa de intensidade da pobreza aumente em média 1.04%, *ceteris paribus*.

$\hat{\beta}_4 = 3.948$: Se (DDR80/20) o Rácio S80/S20 aumentar um valor, estima-se que a Taxa de intensidade da pobreza aumente em média 3.95%, *ceteris paribus*.

Resultados do Segundo Modelo:

Pode concluir-se que ao se alterar, neste Segundo Modelo, a variável dependente "proporção de pessoas em risco de pobreza ou exclusão social", para a variável dependente "Taxa de Intensidade da Pobreza" obtém-se um modelo final completamente diferente daquele obtido na experiência do Primeiro Modelo. O que se mostra, de facto, interessante, sendo que em ambos os modelos a questão base analisada é: evolução da pobreza em Portugal.

No Primeiro Modelo analisou-se a proporção da população residente em risco de

pobreza ou exclusão social, ou seja, a proporção de indivíduos que vivem em Portugal em risco de pobreza e/ou em situação de privação material severa e/ou a viver em agregados com intensidade laboral per capita muito reduzida. E, como tal, obtiveram-se as Taxa de privação material e o Rácio S90/S10 como sendo as variáveis que mostram explicar da melhor forma a variação da variável dependente.

Neste Segundo Modelo, ao ser analisada a intensidade da pobreza está a ser avaliada a variação desta intensidade nos seus diversos níveis, tendo-se constatado que a carga das despesas habitacionais, o desemprego registado, o Índice de dependência total e o Rácio S80/S20 são as variáveis com maior peso para explicar a variação do comportamento da variável Taxa de Intensidade da Pobreza.

Outra conclusão que se pode retirar é a de que apesar dos valores do desemprego registado (variável "DR") fazerem a percentagem da intensidade da pobreza variar muito pouco, esta mostra-se ser uma variável bastante significativa para este modelo. Resultado espectável em face dos estudos disponibilizados na fase de revisão da literatura feita no início do estágio.

Dos quais se salienta nomeadamente: "(…) praticamente um terço dos pobres (quase 33%, ou seja, quase um em cada três) são trabalhadores. (…) percebe-se que mais de metade das pessoas em situação de pobreza trabalha, o que significa que ter um emprego seguro não é suficiente para sair de uma situação de pobreza.". [20] Como afirma o estudo "Pobreza em Portugal – Trajetos e Quotidianos"[21]: "(…) as pessoas em situação de pobreza em Portugal: 32,9% são trabalhadores, 27,5% serão reformados, 26,6% são precários e 13% são desempregados", repara-se que a percentagem de pessoas em situação de pobreza que trabalha[22] é muito elevada, contudo, sendo que maior parte destes tem vínculos laborais sem termo e auferem o salário mínimo que têm de dividir com a família - em muitos casos com uma família numerosa - não conseguem sair desta situação, como é explicado pelo Prof. Fernando Diogo, coordenador do estudo acima referido.

---

[20] (Porfírio, J. (2021, 12 abril). "Um quinto da população portuguesa é pobre. Um em cada três pobres tem emprego estável". *Observador*. https://observador.pt/2021/04/12/a-pobreza-em-portugal-quase-20- das-pessoas-sao-pobres-e-um-em-cada-tres-pobres-tem-emprego-estavel/)

[21] Diogo, F., Farinha Rodrigues, C., Palos, A.C., Pereira, E., Bessa, F., Trevisan, G., Fernandes, L., Silva, O., Perista, P., Amaro, I. "A Pobreza em Portugal: Trajetos e Quotidianos". (2021, abril). Fundação Francisco Manuel dos Santos. Disponível em: https://www.ffms.pt/pt-pt/estudos/pobreza-em-portugal-trajetos-e-quotidianos

[22] Esta percentagem não esta incluída na nossa variável de estudo: Desemprego Registado ("DR").

Efetivamente fará sentido a variável "DR" se mostrar significativa para o modelo visto que 43.4% dos desempregados em Portugal estão em situação de pobreza (segundo dados do Inquérito às Condições de Vida e Rendimento (ICOR 2021) – INE), o que significa que são o grupo onde a Taxa de pobreza é mais elevada. Contudo, a baixa variação que se verificou na análise deste modelo permite concluir que sair de uma situação de desemprego não significa sair de uma situação de pobreza. Como referido acima, é importante compreender que as pessoas em situação de pobreza não são todas iguais.

Relativamente à variável independente Índice de Dependência Total (IDT), esta permaneceu no modelo final, mostrando-se bastante significativa estatisticamente, possuindo o menor *p-value* dentro das variáveis do modelo (ver Figura 15), como seria expectável, uma vez que esta diz respeito à relação entre a população jovem e idosa e a população em idade ativa. A evolução demográfica em Portugal tem vindo a caracterizar-se por um gradual aumento do peso dos grupos etários seniores e uma redução do peso relativo da população em idade ativa. As projeções para a evolução da população em Portugal é uma questão muito discutida atualmente.

Nesta matéria importa, ainda, destacar a "dimensão familiar" da pobreza, uma vez que muitos dos considerados em situação de pobreza o são ou porque não têm rendimentos, ou porque estes são irregulares, ou baixos, e têm de os partilhar. O Prof. Fernando Diogo, no estudo "A Pobreza em Portugal: Trajetos e Quotidianos" referido anteriormente, alerta para duas tipologias de famílias com taxas de pobreza acima da média global: famílias monoparentais e famílias onde existem dois adultos com três ou mais crianças. Afirmando que "os agregados onde existem crianças são aqueles em que a taxa de pobreza é mais elevada". Fazendo assim todo o sentido que esta variável se mantenha incluída no modelo.

Por fim, é curioso notar que para esta variável dependente o melhor rácio - indicador de desigualdade na distribuição do rendimento - para explicar a variação da Taxa de Intensidade de Pobreza é o Rácio S80/S20. Ao contrário do que se verificou no Primeiro Modelo, neste modelo é preferido o indicador que observa um maior número de indivíduos nos seus extremos para ser analisada a intensidade da pobreza. Este rácio permite analisar a desigualdade entre estes extremos da distribuição. E, tal como é referenciado por Carlos Farinha Rodrigues[23], o panorama em Portugal é marcado por

[23] Carlos Farinha Rodrigues, autor de vários estudos sobre a pobreza e a desigualdade em Portugal, como também sobre a eficácia de políticas sociais, é professor associado no Instituto Superior de Economia e

"um forte agravamento da desigualdade assente no afastamento entre os extremos da distribuição e uma certa estabilização das assimetrias existentes na sua parte central. Dado que o rendimento médio dos indivíduos situados na parte superior da distribuição não cresceu, o agravamento da desigualdade é, pois, indissociável da forte contração dos rendimentos mais baixos", o que leva a crer que faz todo o sentido incluir uma variável como esta no modelo de explicação da Taxa de Intensidade de Pobreza.

Verificação dos pressupostos do modelo final:

- Normalidade:

Este pressuposto foi analisado analiticamente, através do teste de Kolmogorov-Smirnov.

$KS_{obs} = 0.16825$

p-value= $0.2264 > \alpha = 0.05$

Decisão: Não rejeitar a hipótese nula.

Conclui-se que, com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que os resíduos não seguem uma Distribuição Normal.

- Média nula:

Pelo *software* R confirma-se que a média dos resíduos é aproximadamente zero.

- Homocedasticidade

Para uma análise analítica deste pressuposto realizou-se o Teste de Breusch-Pagan:

$BP_{obs} = 3.8386$

p-value= $0.4283 > \alpha = 0.05$

Decisão: Não se rejeita H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, conclui-se que não existe evidência estatística de que a variância dos resíduos não é homogénea, o pressuposto está validado.

Gestão (ISEG), Universidade de Lisboa e consultor no Instituto Nacional de Estatística (INE) na área de estatísticas dos agregados familiares. Menção ao seu estudo: Rodrigues, C.F et al. *Desigualdade do Rendimento e Pobreza em Portugal*, Lisboa, FFMS.

- Independência dos resíduos

Para verificar o pressuposto da independência dos resíduos realizou-se o Teste de Durbin-Watson:

DW_{obs}= 2.417

p-value= 0.883 > α = 0.05

Decisão: Não rejeitar H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que existe autocorrelação.

- Ausência de Multicolinearidade

Todos os VIF < 10

Conclui-se que todos os pressupostos são válidos, é possível utilizar o modelo para fazer previsão.

Pesquisa de observações influentes:

```
> dcook[dcook>1]
named numeric(0)
> lev[lev>0.5]
named numeric(0)
```

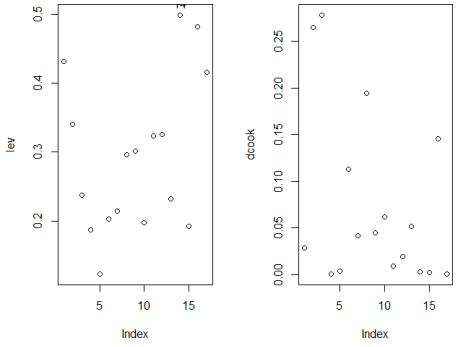

Figura 16 - Valores de Leverage e Distância de Cook

Na Figura 16 observa-se que nenhuma das observações será considerada um *outlier*, visto nenhuma mostrar ter um valor de Leverage superior a 0.5. Da mesma forma que nenhuma das observações será considerada uma observação influente, visto que nenhuma mostra ter uma distância de Cook superior a 1, pela representação gráfica é até mesmo comprovado que todas têm uma distância de Cook muito inferior a 1. Assim sendo, o ajuste do modelo não é afetado.

Previsão:

De seguida, utiliza-se o modelo final obtido para estimar/prever um intervalo de confiança para a variável dependente com base nos valores das variáveis independentes. Recorrendo ao *software* estatístico R.

Tal como descrito no Primeiro Modelo, começa-se por observar o intervalo de variação das variáveis independentes, e, a partir daí, alteram-se os valores destas variáveis observando a variação da variável Taxa de intensidade da pobreza ("TIP") noutros cenários/contextos.

Tabela 7: Previsão da variável dependente TIP

Ano	Valor pontual estimado para "TIP"	Intervalo de Confiança (95%) individual de predição de nova observação	Valor real de "TIP"	"CDH"	"DR"	"IDT"	"DDR80/20"
2020	26.8%	(24.41 ; 29.20)	27.1%	10.4*	402254*	55.9*	5.7*

*= Valores reais das variáveis. Retirados do Portal do INE

Como é possível observar na Tabela 7, no ano de 2020, o valor real da variável dependente está contido no intervalo de predição calculado, pelo que se pode concluir que o modelo estará a funcionar corretamente.

Tabela 8: Cenário

Ano	Valor pontual estimado para "TIP"	Intervalo de Confiança (95%) individual de predição de nova observação	Valor real de "TIP"	"CDH"	"DR"	"IDT"	"DDR80/20"
2021	24.2%	(21.76 ; 26.54)	21.7%	10.5*	347959*	56.1	5.1*
2022	23.5%	(21.07 ; 25.91)		10.2*	307005*	56.3	5
2023	24.4%	(22.02 ; 26.83)		10.8	307005	56.9	5
2024	24.3%	(21.83 ; 26.81)		10.2	307005	57.1	5

*= Valores reais das variáveis. Retirados do Portal do INE

Para o ano de 2021 – Tabela 8 – começou-se por se observar as estatísticas descritivas das variáveis independentes e para a variável Índice de dependência total ("IDT") teve-se em conta o seu aumento ao longo dos anos (aumento este que é possível se observar de imediato na base de dados). Assim sendo, decidiu-se aumentar os valores desta. Obteve-se assim, para o valor da variável "TIP", um intervalo de confiança individual de predição que por pouco não inclui o valor real da variável. Contudo, a

enorme descida no valor real de "TIP" de 2020 para 2021 é representada pelo modelo.

Em relação ao cenário para o ano de 2022, com base na variação das variáveis, decidiu-se manter a subida do Índice de dependência total ("IDT"), como referido anteriormente. E para a variável "DDR80/20" que tem apresentado uma diminuição no seu valor todos os anos (com exceção de 2020), decidiu-se diminuir ligeiramente. O que resultou numa descida do valor pontual estimado para a variável dependente e do respetivo intervalo de predição, o que traduz/revela um cenário otimista de continuação da descida desta Taxa de intensidade da pobreza.

No cenário do ano de 2023, para o qual ainda não se encontram disponíveis quaisquer dados, tentou-se criar um cenário realista, onde o valor da variável carga das despesas em habitação ("CDH") apresentasse um ligeiro aumento - com base na conjuntura atual vivida; as variáveis inseridas no modelo: desemprego registado ("DR") e "DDR80/20" se mantivessem nos mesmos valores; e o valor da variável "IDT" continuasse a aumentar. O que, como é possível observar na Tabela 8, resulta num aumento significativo do valor pontual estimado para "TIP" para este ano.

Para o cenário do ano de 2024 decidiu-se fazer a experiência de diminuir novamente o valor da variável "CDH", ou seja, como se este se tivesse mantido desde 2022, ou decrescido em 2024, deixando, de qualquer das formas, o padrão de subida do Índice de dependência total ("IDT"). O que resultaria numa passagem do valor estimado para a Taxa de intensidade da pobreza de 24.4% para 24.3%, em relação ao cenário hipotético criado para o ano anterior.

Tabela 9: Dois cenários

Ano	Valor pontual estimado para "TIP"	Intervalo de Confiança (95%) individual de predição de nova observação	Valor real de "TIP"	"CDH"	"DR"	"IDT"	"DDR80/20"
2028	24.7%	(22.14 ; 27.23)		10.1	307005	57.5	5
2028	23.9%	(21.32 ; 26.47)		10.1	307005	57.5	4.8

Por fim, apresenta-se na Tabela 9 a experiência de dois cenários para daqui a 5 anos. Onde, em ambos, o panorama do Índice de dependência total, renovação da população e falta de população em idade ativa piora ao longo do tempo. Nestes casos, o

valor da variável “IDT” aumentaria 0.10 pontos percentuais por cada ano. Apresentando-se todas as variáveis com os mesmos valores em ambos os cenários, com exceção da variável “DDR80/20”, assim, confirma-se que é possível reduzir a Taxa de intensidade da pobreza (“TIP”) mesmo que a tendência seja que o quadro da variável “IDT” piore nos próximos anos. Através da descida do indicador de desigualdade na distribuição do rendimento, Rácio S80/S20, observa-se a descida mais significativa do valor estimado para a variável dependente, este mostrou ser o resultado mais significativo de todas as alterações de valores das restantes variáveis independentes que foram experimentadas nesta experiência.

Concluindo, é possível reduzir a Taxa de Intensidade de Pobreza, mesmo que o panorama no Índice de dependência total seja pior.

O Terceiro e Quarto modelos do estudo incluíram, para além dos dados do INE, os elementos da Carta Social.

Após ponderação as variáveis foram divididas em duas bases de dados. Uma diz respeito à frequência destas respostas (número de utentes), a sua utilização – Excel "dadosgepfreq**"**. Outra diz respeito à capacidade destas respostas (número de lugares), a sua cobertura – Excel "dadosgepcapacidade".

6.3 Terceiro Modelo:

Neste modelo foi escolhida como variável dependente a "Proporção da população residente em risco de pobreza ou exclusão social" = PPRPES, mencionada nos modelos anteriores. E como variáveis independentes foram selecionadas algumas das respostas sociais integrantes da Rede de Serviços e Equipamentos Sociais (RSES) no grupo-alvo Família e Comunidade, disponível na Carta Social.

Objetivo do Terceiro Modelo: Avaliar se o número de utentes nas respostas sociais em estudo (variáveis independentes que se encontram na Tabela 10) explica a variação da variável dependente PPRPES.

Tabela 10: Variáveis do modelo

Variáveis Independentes	**Abreviaturas**	**Unidade**
Centro de Alojamento Temporário	CAT	Nº
Comunidade de Inserção	CI	Nº
Refeitório/Cantina Social	R	Nº
Serviço de Atendimento e Acompanhamento social	SAAS	Nº
Ajuda Alimentar	AA	Nº

6.3.1 Estatística descritiva das variáveis RSES Família e Comunidade:

Tabela 11: Análise das Respostas Sociais

Resposta Social	Nº utentes Frequência	Nº lugares Capacidade	Taxa de utilização %
Serviço de Atendimento e Acompanhamento Social	91602	116617	78,55%
Refeitório/Cantina Social	4139	7051	58,70%
Comunidade de Inserção	5250	6272	83,71%
Centro de Alojamento Temporário	1643	1934	84,95%
Ajuda alimentar	94079	100518	93,60%

Fonte: Carta Social, GEP-MTSSS, 2021

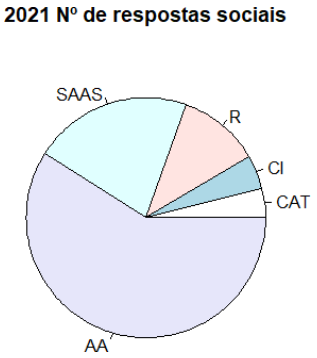

Figura 17 - Nº de respostas Sociais, Portugal - 2021

Como é possível observar no gráfico da Figura 17, a resposta social Ajuda Alimentar (AA) é aquela que tem maior cobertura de respostas a nível nacional. Pode-se afirmar que existem mais equipamentos com o propósito de ajuda alimentar, seguido de equipamentos com a funcionalidade de Serviço de Atendimento e Acompanhamento Social (SAAS).

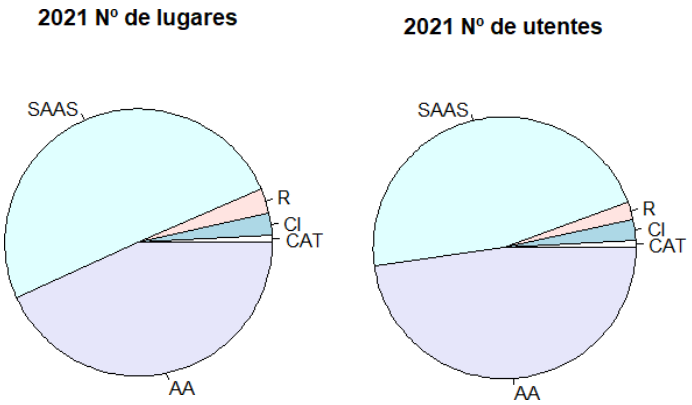

Figura 18 - Capacidade e Frequência das respostas sociais, Portugal - 2021

Ao analisar os gráficos presentes na Figura 18, conclui-se que a Ajuda Alimentar (AA), para além de ter uma grande cobertura de lugares a nível nacional, tem também taxa de utilização de 93.60%, apesar de ter uma capacidade um pouco inferior à do Serviço de Atendimento e Acompanhamento Social (SAAS). É clara a procura destas respostas sociais por parte da população portuguesa.

Pode também notar-se que as restantes respostas sociais: Comunidade de Inserção (CI) e Centro de Alojamento Temporário (CAT) não alteram muito os seus valores entre o gráfico referente ao número de lugares existentes e o gráfico referente ao número de utentes que estão inscritos nas mesmas, o que significa que estes estão a ser bastante utilizados pela população para a qual se destinam. À exceção da resposta social Refeitório/Cantina Social (R), que mostra um decréscimo da área em relação ao número de pessoas que utilizam esta resposta, o que, do ponto de vista da autora, pode acontecer devido à finalidade desta ser muito idêntica ao propósito da Ajuda Alimentar (AA). Ou, talvez, pela eventual preferência em recorrer à Ajuda Alimentar em vez da utilização dos Refeitórios/Cantinas, por parte dos indivíduos e famílias nestas situações de vulnerabilidade, por questões de vergonha e aceitação social - a própria administração da Santa Casa da Misericórdia de Lisboa refere que a frequência desta resposta social é sempre maior no último trimestre do ano, perto do Natal, quando existe um maior sentimento de solidariedade entre a população, e menos vergonha[24]. Ou, numa outra possibilidade, pelo desconhecimento da localização ou difícil acesso/transporte até aos locais onde estes equipamentos se encontram. Visto que, ao contrário da Ajuda Alimentar (que distribui alimentos através de entidades), os Refeitórios ou Cantinas são locais fixos, destinados ao fornecimento de refeições, a indivíduos e famílias em situação de vulnerabilidade socioeconómica.

Em 2021 o Refeitório, ou Cantina Social, tem capacidade para receber quase mais 50% de utilizadores do que o que tem recebido. Assim sendo, caso esta variável se inclua no modelo final apto para a realização de previsão poderá ser feita a experiência de alterar os seus valores e observar se tal influenciará a variável dependente PPRPES, é esperado que tal aconteça.

6.3.2 Modelação:

No Terceiro Modelo a equação a ajustar é a seguinte:

Modelo a ajustar:

$$PPRPES_i = \beta_0 + \beta_1 CAT + \beta_2 CI + \beta_3 R + \beta_4 SAAS + \beta_5 AA + \varepsilon_i$$

Verificação dos Pressupostos ao modelo inicial:

Os pressupostos encontram-se todos válidos.

Escolha do melhor modelo:

Através dos métodos estatísticos de seleção de variáveis: *Backward*, *Forward* e *Stepwise*, obteve-se a mesma seleção de variáveis para o modelo final.

```
Step:  AIC=7.28
PPRPES ~ CAT + R

       Df Sum of Sq    RSS     AIC
<none>              11.371  7.2848
+ CI    1     1.412  9.959  7.9585
+ SAAS  1     0.882 10.489  8.4770
+ AA    1     0.009 11.362  9.2773
- CAT   1    24.952 36.323 16.8986
- R     1    43.721 55.091 21.0641

Call:
lm(formula = PPRPES ~ CAT + R, data = dados)

Coefficients:
(Intercept)          CAT            R
  -2.451219     0.010518     0.001714
```

Figura 19 - "Melhor modelo" pelo método estatístico "Backward"

[24] Leitura: *Observador*. "Cantinas sociais em Lisboa sem procura excecional mantêm número de refeições". Disponível em: https://observador.pt/

Modelo final a ajustar:

$PPRPES_i = \beta_0 + \beta_1 CAT + \beta_2 R + \varepsilon_i$

Modelo final ajustado:

$\hat{PPRPES}i = -2.451 + 0.011\ CAT + 0.002\ R$

```
> summary(modeloterceiro2)

Call:
lm(formula = PPRPES ~ CAT + R)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3547 -0.7866 -0.1280  0.3795  2.4671

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.4512188  5.4701700  -0.448  0.66762
CAT          0.0105176  0.0026836   3.919  0.00575 **
R            0.0017139  0.0003304   5.188  0.00127 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.275 on 7 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8202,    Adjusted R-squared:  0.7688
F-statistic: 15.97 on 2 and 7 DF,  p-value: 0.002465
```

Figura 20 - Modelo 3 - Modelo final

Qualidade do ajustamento:

R_a^2 = 0,769. Coeficiente de determinação varia entre 0 e 1. Qualidade boa. 77% da variabilidade total da % de PPRPES é explicada pela equação ajustada.

Significância global do modelo:

Teste F da ANOVA:

H_0: $\beta_1 = \beta_2 = 0$

H_1:] $\beta_i \neq 0$

i = 1,2

F_{obs} = 15,97

p-value= 0,00247 < α = 0,05 Decisão: Rejeitar H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, existe evidência estatística de que o modelo é globalmente significativo.

Testar a significância individual:

Para estudar a significância individual utilizamos os testes t:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

i = 1,2

	T_{obs}	*p-value*
β_1	3,919	0,00575 < α
β_2	5,188	0,00127 < α

Decisão: Rejeitar H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, existe evidência estatística de que ambas as variáveis são significativas individualmente.

Interpretação dos coeficientes ajustados:

$\hat{\beta}_0 = -2{,}451$: ordenada na origem, só faria sentido interpretar se as restantes variáveis assumissem o valor zero.

$\hat{\beta}_1 = 0{,}011$: Se a frequência dos Centros de Alojamento Temporários (CAT) aumentar mais uma pessoa, estima-se que a % de PPRPES aumente em média 0,011%, *ceteris paribus*.

$\hat{\beta}_2 = 0{,}002$: Se a frequência dos Refeitórios ou Cantinas Sociais (R) aumentar mais um individuo, estima-se que a % de PPRPES aumente em média 0,002%, *ceteris paribus.*

Verificação dos pressupostos do modelo final:

- Normalidade:

Analisando graficamente, na Figura 21 observa-se que a maioria dos pontos, isto é, quantis das observações, se encontram próximos da linha que representa os quantis da Normal, isto é, os resultados esperados se os dados seguissem uma distribuição normal.

Pode-se concluir que os resíduos apresentam, pelo menos aproximadamente, uma Distribuição Normal.

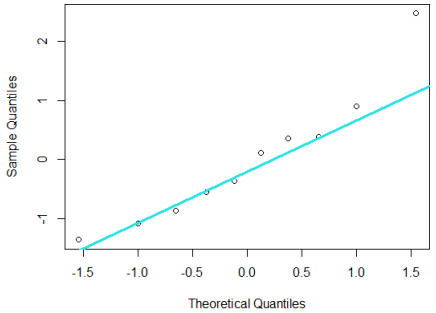

Figura 21 - Q-Q Plot Normalidade

```
> lillie.test(resid(modeloterceiro2))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(modeloterceiro2)
D = 0.16549, p-value = 0.611

> shapiro.test(resid(modeloterceiro2))

        Shapiro-Wilk normality test

data:  resid(modeloterceiro2)
W = 0.92444, p-value = 0.3955
```

Para confirmar este pressuposto, realizou-se uma análise analítica. Tanto no teste de Kolmogorov-Smirnov como no teste de Shapiro-Wilk, vemos que os *p-values* são superiores a $\alpha = 0.05$, o que nos leva a não rejeitar a hipótese nula. Conclui-se que com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que os resíduos não seguem uma Distribuição Normal.

- Média nula:

Pelo *software* R confirma-se que a média dos resíduos é aproximadamente zero.

- Homocedasticidade:

Para verificar a Homocedasticidade optou-se pela realização do teste de Breusch-Pagan:

```
> bptest(modeloterceiro2)

        studentized Breusch-Pagan test

data:  modeloterceiro2
BP = 0.099561, df = 2, p-value = 0.9514
```

$BP_{obs} = 0.09956$

p-value$= 0.951 > \alpha = 0.05$

Decisão: Não rejeito H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, conclui-se que não existe evidência estatística de que a variância dos resíduos não é homogénea, o pressuposto está validado.

- Independência:

Para verificar o pressuposto da independência dos resíduos realizou-se o Teste de Durbin-Watson:

```
> dwtest(modeloterceiro2, alternative = "two.sided")

        Durbin-Watson test

data:  modeloterceiro2
DW = 1.3938, p-value = 0.04759
alternative hypothesis: true autocorrelation is not 0
```

DW_{obs}= 1.3938

p-value= 0.048 < α = 0.05

Decisão: Não rejeitar H0

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que existe autocorrelação.

- Ausência de Multicolinariedade:

```
     CAT        R
1.140714 1.140714
```

Ambos os VIF < 10.

Conclui-se que todos os pressupostos são válidos, é possível utilizar o modelo para fazer previsão.

Resultados do Terceiro Modelo:

Com base neste modelo aprende-se que, como indicadores da variação da proporção de pessoas em risco de pobreza ou exclusão social em Portugal, serão estatisticamente significativas as variáveis "CAT" (Centro de Alojamento Temporário) e "R" (Refeitório), em relação ao número de utentes que frequenta cada uma delas. O que não fora previsto imediatamente no início do estudo, sendo que estas respostas sociais não eram as que mais se destacavam no apoio a indivíduos em situação de vulnerabilidade socioeconómica.

Contudo, ao se analisar as respostas sociais (Tabela 11) este resultado era esperado, visto que a frequência em comparação à capacidade nos Refeitórios, ou Cantinas Sociais, é menor do que em qualquer outro equipamento social mencionado no estudo. Em várias respostas que possuem esta finalidade os valores têm-se mantido os mesmos - como é exemplo a Santa Casa da Misericórdia de Lisboa, que afirma no artigo mencionado acima que "a instituição não tem recebido mais pedidos de apoio ao nível da Cantina Social, que tem mantido o número de utentes" - fará sentido que a alteração dos valores da frequência nos Refeitórios/Cantinas seja significativa para a explicação da variação da variável dependente. Isto é, seria lógico que ao aumentar o número de indivíduos que frequentam os Refeitórios (R) se observe também um aumento na proporção de população em risco de pobreza ou exclusão social (PPRPES).

A variável Centros de Alojamento Temporário (CAT), em relação ao número de utentes, mostra ser outra variável significativa para o modelo. Esta resposta social que tem como objetivos: o apoio em alojamento; o desenvolvimento de atividades e serviços que se destinem à promoção e integração social do indivíduo; e a promoção da pessoa e combate à exclusão social, tem uma elevada taxa de utilização (aproximadamente 85%), e sendo que não existem muitas a nível nacional, é interessante que esta seja uma das variáveis independentes que explica da melhor forma a variação da variável dependente.

Concluindo, ao observar o panorama nestas duas respostas, nomeadamente, a frequência (número de utentes), é possível compreender a variação do comportamento da proporção de população residente em Portugal em risco de pobreza ou exclusão social.

Previsão:

Com o *software* estatístico R foram realizados cálculos de previsão através do modelo final obtido, cujos pressupostos se encontram todos validados.

Nesta situação existe interesse em fazer previsão intervalar individual, isto é, prever possíveis valores para a variável dependente referido a outro período ou a outro contexto.

Quando se começou a modelar só se encontravam disponíveis valores da variável dependente "PPRPES" até ao ano de 2020. Entretanto foram disponibilizados valores para 2021 e 2022, pelo que se utilizarão estes valores para confirmar se o modelo está a funcionar corretamente, isto é, se as previsões espelham a realidade.

Em relação aos dados facultados pelo GEP, os valores mais recentes disponíveis dizem respeito ao número de utentes nestas respostas sociais em 2021. Tais foram: 1643 utentes nos Centros de Alojamento Temporários (CAT), e 4139 utilizadores do Refeitório/Cantina Social (R). Assim sendo, como se pode observar na Tabela 12, foi calculada uma estimativa do valor de "PPRPES" para 2021 que poderá ser comparada com o valor real obtido posteriormente.

Tabela 12: Previsão da variável dependente PPRPES

Ano	Valor pontual estimado para "PPRPES"	Intervalo de Confiança (95%) individual de predição de nova observação	Valor real de "PPRPES"	"R"	"CAT"
2020	21.2%	(17.71 ; 24.60)	19.8%	4016 *	1590 *
2021	21.9 %	(18.61 ; 25.24)	22.4%	4139 *	1643 *
2022	22.1%	(18.80 ; 25.40)	20.1%	4239	1643

*= Estes dados são valores reais do número de utentes nestas respostas sociais. Dados disponibilizados pelo GEP.

Conclui-se que foi feita uma boa modelação. Ao observarmos a Tabela 12 confirma-se que foi feita uma boa previsão pois o valor real está incluído/contido dentro dos intervalos de previsão obtidos. Os valores pontuais estimados, tanto para 2021 como para 2022, não se afastam muito dos valores reais que se obtiveram posteriormente.

Sendo que a informação relativa à Carta Social só se encontra disponível com referência ao mês de dezembro de cada ano, ainda não é possível ter acesso a valores reais para o ano de 2023. Posto isto, decidiu-se realizar várias experiências de previsão de estimativas das quais três cenários são apresentados nas tabelas abaixo.

Tabela 13: Cenário 1

Ano	Valor pontual "PPRPES"	Intervalo de Confiança (95%) nova observação	Valor real "PPRPES"		
2022	22.1%	(18.80 ; 25.40)	20.1%	4239	1643
2023	22.3%	(18.98 ; 25.55)		4339	1643
2024	22.4%	(19.17 ; 25.70)		4439	1643

Os valores nos Refeitórios Sociais mostram que a frequência destes estava em significativo decréscimo todos os anos desde 2014, contudo, no ano de 2020 a situação mudou e verificou-se um aumento de 244 utentes e no ano de 2021 outro aumento de 123 utentes.

Os valores nos Centros de Alojamento Temporário têm tido muito pouca variabilidade. Este conjunto de dados apresenta-se em torno dos mesmos valores ao longo dos últimos anos 10 anos.

Na Tabela 13 estamos perante um cenário pessimista, onde o número de utentes na variável "R" (hipoteticamente) aumentaria novamente em 2022, em 2023 e em 2024 – um valor de 100 utentes por ano. A variável "CAT" mantinha sempre o mesmo valor, algo não muito possível de acontecer, mas, sendo que os valores desta variável têm tido poucas alterações, não torna o cenário irrealista, permitindo focar o estudo na resposta social "R".

É ainda possível confirmar que neste cenário pessimista a variável dependente aumentaria 0.1 pontos percentuais por ano com o aumento da frequência (número de utentes) na resposta social Refeitório/Cantina Social.

Tabela 14: Cenário 2

Ano	Valor pontual estimado para "PPRPES"	Intervalo de Confiança individual de predição de nova observação	Valor real de "PPRPES"	"R"	"CAT"
2020	21.2%	(17.71 ; 24.60)	19.8%	4016 *	1590 *
2021	21.9 %	(18.61 ; 25.24)	22.4%	4139 *	1643 *
2022	21.8%	(18.42 ; 25.09)	20.1%	4039	1643

*= Estes dados são valores reais do número de utentes nestas respostas sociais. Dados disponibilizados pelo GEP.

Pelo contrário, no Cenário 2 (Tabela 14) experimentou-se em 2022 diminuir o número de utentes na variável "R" e manter o número de utentes na variável "CAT", o que resultou num intervalo de predição onde o valor real de "PPRPES" se inclui. Contudo, o valor pontual estimado para a variável dependente mostrou-se mais elevado do que o valor real.

Neste seguimento, experimentou-se diminuir o número de utentes em ambas as variáveis independentes, criando um cenário mais otimista (Cenário 3 – Tabela 15) onde a estimativa para "PPRPES" e o seu valor real para 2022 se aproximam um pouco mais.

Ainda nesta Tabela 15 apresenta-se a experiência, para o ano de 2023, de diminuir o número de utentes na variável "R" (em 100 utentes em relação aos dados reais de 2021, isto é, 50 utentes em 2022 e 50 utentes em 2023) e aumentar o número de utentes na variável "CAT" (em 50 utentes em relação aos dados reais de 2021, isto é, 25 utentes por ano), o que resultou num aumento quase de 1% do valor pontual estimado para a variável "PPRPES".

Por fim, apresenta-se também na Tabela 15 uma experiência para o ano de 2024 onde, desta vez, o valor de "R" se mantivesse ao longo dos anos e o valor de "CAT" descesse em 20 utentes por ano. Repara-se que se obtém um valor estimado da variável dependente e um intervalo de predição muito idênticos/semelhantes ao que se obteve quando se experimentou o cenário otimista de diminuir o número de utentes em ambas as variáveis independentes. Ou seja: Alterar os valores da variável "CAT" parece influenciar mais os valores estimados para a variável dependente (do que alterar os valores da variável "R")

Tabela 15: Cenário 3

Ano	Valor pontual estimado para "PPRPES"	Intervalo de Confiança individual de predição de nova observação	Valor real de "PPRPES"	"R"	"CAT"
2022	21.3%	(17.88 ; 24.72)	20.1%	4039	1600
2023	22.2%	(19.01 ; 25.55)		4039	1693
2024	21.3%	(17.86 ; 24.72)		4139	1583

6.4 Quarto Modelo:

Neste modelo foram utilizadas as mesmas respostas sociais que no Terceiro Modelo. Contudo, neste modelo foram utilizados os dados referentes à capacidade destas respostas sociais (número de lugares), a sua cobertura – Dados: Excel "dadosgepcapacidade".

Objetivo do Quarto Modelo: Avaliar se o número de lugares (capacidade) nas respostas sociais em estudo (variáveis independentes que se encontram acima, na Tabela 11) explica a variação da variável dependente PPRPES.

6.4.1 Modelação:

No Quarto Modelo a equação a ajustar é a seguinte:

Modelo a ajustar:

$PPRPES_i = \beta_0 + \beta_1 CAT + \beta_2 CI + \beta_3 R + \beta_4 SAAS + \beta_5 AA + \varepsilon_i$

Verificação dos Pressupostos ao modelo inicial:

Os pressupostos encontram-se todos válidos.

Escolha do melhor modelo:

Através dos métodos estatísticos de seleção de variáveis: *Backward*, *Forward* e *Stepwise*, obteve-se a mesma seleção de variáveis para o modelo final.

```
Step:  AIC=-1.63
PPRPES ~ CI + SAAS + AA

       Df Sum of Sq     RSS     AIC
<none>              2.3993 -1.6341
+ CAT   1    0.1001  2.2992  0.0251
+ R     1    0.0006  2.3987  0.3639
- AA    1    2.0466  4.4459  1.3004
- SAAS  1    5.7131  8.1124  6.1116
- CI    1   17.5078 19.9072 13.2931

Call:
lm(formula = PPRPES ~ CI + SAAS + AA, data = dados)

Coefficients:
(Intercept)           CI         SAAS           AA
  2.764e+01    9.525e-04   -1.065e-04   -2.739e-05
```

Figura 22 - "Melhor modelo" pelo método estatístico "Stepwise"

Modelo final a ajustar:

$PPRPES_i = \beta_0 + \beta_1 CI + \beta_2 SAAS + \beta_3 AA + \varepsilon_i$

Modelo final ajustado:

$\hat{PPRPES}i = 27.64 + 0.0009525\,CI - 0.0001065\,SAAS - 0.00002739\,AA$

```
> summary(modeloquatro2)

Call:
lm(formula = PPRPES ~ CI + SAAS + AA)

Residuals:
      3       4       5       6       7       8       9      10
 0.3582  0.3762 -0.7316  0.5913 -0.4922 -0.5901  0.7613 -0.2731

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.764e+01  2.634e+00  10.494 0.000466 ***
CI           9.525e-04  1.763e-04   5.403 0.005682 **
SAAS        -1.065e-04  3.452e-05  -3.086 0.036710 *
AA          -2.739e-05  1.483e-05  -1.847 0.138445
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7745 on 4 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.9613,    Adjusted R-squared:  0.9323
F-statistic: 33.12 on 3 and 4 DF,  p-value: 0.002772
```

Figura 23 - Modelo 4 - Modelo final

Qualidade do ajustamento:

R_a^2 = 0.9323. Coeficiente de determinação varia entre 0 e 1. Qualidade muito boa. 93% da variabilidade total da % de PPRPES é explicada pela equação ajustada.

Significância global do modelo:

Teste F da ANOVA:

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_1:] $\beta_i \neq 0$

i = 1,2,3

F_{obs} = 33.12

p-value= 0.00277 < α = 0.05 Decisão: Rejeitar H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, existe evidência estatística de que o modelo é globalmente significativo.

Testar a significância individual:

Para estudar a significância individual utilizamos os testes t:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$
i = 1,2,3

	T_{obs}	*p-value*
β_1	5.403	$0.0057 < \alpha$
β_2	-3.086	$0.0367 < \alpha$
β_3	-1.847	$0.1384 > \alpha$

Os métodos *backward, forward* e *stepwise* escolheram um modelo cujas variáveis independentes serão: Comunidade de Inserção ("CI"), Serviço de Atendimento e Acompanhamento Social ("SAAS") e Ajuda Alimentar ("AA").

Relativamente a testar a significância individual, com base nestes dados amostrais, e ao nível de significância de 5%, existe evidência estatística de que as variáveis "CI" e "SAAS" são significativas individualmente.

Já a variável "AA", com um *p-value* superior a alfa, não será significativa individualmente. Contudo, visto que esta fora considerada como uma boa candidata para o modelo pelos métodos de seleção de variáveis utilizados no R – Ver Figura 22 -, após reflexão, e confirmação se o sinal das correlações entre cada uma das variáveis independentes e a dependente coincide com o sinal dos coeficientes do modelo, decidiu-se manter a variável "AA" no modelo.

Interpretação dos coeficientes ajustados:

$\hat{\beta}_0 = 27.64$: ordenada na origem, só faria sentido interpretar se as restantes variáveis assumissem o valor zero.

$\hat{\beta}_1 = 0.0009525$: Se a capacidade das CI aumentar, isto é, se se disponibilizar mais um lugar, estima-se que a PPRPES aumente em média 0.0009525 %, *ceteris paribus.*

$\hat{\beta}_2 = -0.0001065$: Se a capacidade do SAAS aumentar, se se disponibilizar mais um lugar, estima-se que a % de PPRPES diminua em média 0.0001065 %, *ceteris paribus*

$\hat{\beta}_3 = -0.00002739$: Se a capacidade da AA aumentar 1 unidade, estima-se que a % de PPRPES diminua em média 0.00002739 %, *ceteris paribus.*

Verificação dos pressupostos do modelo final:

- Normalidade:

Este pressuposto foi analisado analiticamente, através do teste de Kolmogorov-

Smirnow e do teste de Shapiro-Wilk.

KS_{obs} = 0.22966

p-value= 0.245 > α = 0.05

SW_{obs} = 0.89413

p-value= 0.2556 > α = 0.05

Tanto no teste de Kolmogorov-Smirnov como no teste de Shapiro-Wilk, os *p-values* são superiores a α = 0.05, o que nos leva a não rejeitar a hipótese nula.

Conclui-se que, com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que os resíduos não seguem uma Distribuição Normal.

- Média nula:

Pelo *software* R confirma-se que a média dos resíduos é aproximadamente zero.

- Homocedasticidade:

Para uma análise analítica deste pressuposto realizou-se o Teste de Breusch-Pagan:

BP_{obs} = 1.0775

p-value= 0.7825 > α = 0.05

Decisão: Não rejeito H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, conclui-se que não existe evidência estatística de que a variância dos resíduos não é homogénea, o pressuposto está validado.

- Independência dos resíduos:

Para verificar o pressuposto da independência dos resíduos realizou-se o Teste de Durbin-Watson:

DW_{obs}= 2.9414

p-value= 0.6396 > α = 0.05

Decisão: Não rejeitar H_0.

Conclusão: Com base nestes dados amostrais e ao nível de significância de 5%, não existe evidência estatística de que existe autocorrelação.

- Ausência de Multicolinearidade:

Todos os VIF < 10

Conclui-se que todos os pressupostos são válidos, é possível utilizar o modelo para fazer previsão.

Resultados do Quarto Modelo:

Com base neste modelo aprende-se que, como indicadores da variação da proporção de pessoas em risco de pobreza ou exclusão social em Portugal, serão estatisticamente significativas as variáveis "CI" (Comunidade de Inserção), e "SAAS" (Serviço de Atendimento e Acompanhamento Social), em relação ao número de lugares que cada uma faculta, ou seja, à sua capacidade.

A variável "AA" (Ajuda Alimentar) foi mantida no modelo pois, apesar de não passar no teste de significância individual, é aquela que tem maior cobertura a nível nacional, e a que tem a maior taxa de utilização pela população (aproximadamente 94%), tendo sido apontada pela equipa do GEP como sendo muito importante a sua participação nos estudos. A resposta social Ajuda Alimentar tem como objetivos a distribuição de alimentos, no continente, através do apoio de diversas entidades nomeadamente do Fundo Europeu de Auxílio aos Carenciados contribuindo para a resolução de situações de carência alimentar e falta de dignidade das famílias e, consequentemente, para a diminuição da proporção de população residentes em risco de pobreza ou exclusão social (PPRPES).

A relevância dos lugares disponíveis em ambas as variáveis "CI" e "SAAS" para a explicação da variação da variável dependente do modelo não foi inesperado.

Sendo a "SAAS" a resposta social com a segunda maior cobertura a nível nacional, e tendo como propósito um vasto leque de atributos, tais como: informar, aconselhar e encaminhar para respostas, serviços ou prestações sociais adequadas a cada situação; e assegurar o acompanhamento social do percurso de inserção social, a sua permanência no modelo foi coerente com o esperado no início do estudo.

Por fim, a "CI", variável mais significativa individualmente no modelo, diz respeito a infraestruturas de apoio social com o objetivo de satisfazer as necessidades básicas e o desenvolvimento de capacidades e potencialidades das pessoas e famílias, através do estímulo à participação nas atividades e do apoio na definição do seu projeto de vida, são uma das respostas sociais cuja capacidade tem vindo a crescer nos últimos anos, consequentemente impacta a variação da variável "PPRPES".

Previsão:

Com o *software* estatístico R foram realizados cálculos de previsão através do modelo final obtido, cujos pressupostos se encontram todos validados.

Nesta situação existe interesse em fazer previsão intervalar, isto é, prever possíveis valores para a variável dependente referido a outro período ou a outro contexto.

Os dados mais recentes disponíveis para a realização deste estudo dizem respeito ao número de lugares nas respostas sociais em estudo, ou seja, a sua capacidade, em 2021. Estes valores são os seguintes: 6272 lugares nas Comunidades de Inserção (CI), 100518 como capacidade total na Ajuda Alimentar (AA), e 116617 lugares para o Serviço de Apoio e Acompanhamento Social (SAAS).

Assim, serão utilizados os valores reais destas três variáveis em 2021 para se obter a previsão da "PPRPES" e, de seguida, comparar esta com o valor real que se obteve posteriormente.

Tabela 16: Previsão da variável dependente PPRPES

Ano	Valor pontual estimado para "PPRPES"	Intervalo de Confiança (95%) individual de predição de nova observação	Valor real de "PPRPES"	"CI"	"AA"	"SAAS"
2020	20.1%	(17.22 ; 22.93)	19.8%	6105 *	78614 *	105391 *
2021	18.4%	(12.88 ; 23.99) **	22.4%	6272 *	100518 *	116617 *
2022	16.8%	(12.12 ; 21.57)	20.1%	6372	185518	110617

*= Estes dados são valores reais do número de lugares nestas respostas sociais. Dados disponibilizados pelo GEP.

**= Intervalo de confiança de 99%.

Como é possível observar na Tabela 16, para o ano de 2020, o modelo mostra funcionar bem, o valor real da variável dependente está contido no intervalo de previsão obtido.

Contudo, no ano de 2021, o modelo já não espelhou corretamente a realidade. A variável "PPRPES" apresentou uma subida significativa no seu valor real, de 19.8% em

2020 para 22.4% em 2021. [25] Porém, os cálculos de previsão estimaram uma descida do valor da variável. Para além de que apenas com um intervalo de confiança de 99% é que foi possível obter um intervalo de predição que contivesse este valor real.

Apesar do modelo não ter conseguido acompanhar a subida que ocorreu em 2021 na variável dependente, apresenta-se ainda na Tabela 16 uma experiência para 2022, com base em valores das variáveis independentes próximos da realidade, discutidos com a equipa do estágio.

Ensaiou-se um aumento de 100 lugares nas Comunidades de Inserção ("CI") – o padrão que se tem observado nos últimos anos. Um aumento de 85000 lugares na Ajuda Alimentar ("AA") – um grande aumento, porém, confirmado que poderá ser um valor realista. E diminuíram-se 6000 lugares no Serviço de Atendimento e Acompanhamento Social ("SAAS") – [26]. Desta forma obteve-se uma estimativa mais correta para a variável "PPRPES" cujo valor real apresenta novamente uma descida, estando este contido no intervalo de predição obtido.

Tabela 17: Cenário

Ano	Valor pontual estimado para "PPRPES"	Intervalo de Confiança (95%) individual de predição de nova observação	Valor real de "PPRPES"	"CI"	"AA"	"SAAS"
2023	16.0%	(11.06 ; 20.99)		6472	195518	116617
2024	15.7%	(10.46 ; 21.02)		6572	205518	117617
2025	15.5%	(9.85 ; 21.06)		6672	215518	118617

Como é possível observar na Tabela 17, se a tendência for assistir-se ao aumento do número de lugares em todas as variáveis independentes, neste caso: 100 lugares na "CI"; 10000 lugares na "AA"; e 1000 lugares na "SAAS" por ano, prevê-se um cenário otimista onde a variável dependente diminua de ano em ano. Conclui-se que o

[25] Este foi um período de contexto de pandemia Covid-19. O próprio inquérito foi realizado em condições distintas aos inquéritos anteriores, o que poderá influenciar também os resultados do mesmo.

[26] No âmbito do processo de transferência de competências, em matéria de Serviço de Atendimento e de Acompanhamento Social (SAAS) de pessoas e famílias em situação de vulnerabilidade e exclusão social, para as câmaras municipais (Portaria n.º 63/2021 de 17 de março), em curso, esta resposta social deixará de ser tutelada pela Segurança Social. Neste sentido, a diminuição do número de respostas, capacidade e frequência desta resposta em 2022 deverá ser lida com alguma precaução, uma vez que poderá apenas indicar uma transferência das respostas para as autarquias e não uma redução efetiva.

investimento na capacidade destas respostas sociais pode diminuir a proporção de pessoas residentes em Portugal em risco de pobreza ou exclusão social.

Conclusão

Com este estudo procurou-se apurar se as respostas sociais de apoio à Família e Comunidade têm impacto/correlação com a evolução da pobreza e desigualdades em Portugal, para melhor compreensão, antecipação e adequação dos meios de prevenção do fenómeno da pobreza.

A análise estatística dos dados apresentados, com aplicação dos métodos escolhidos, permitiu concluir que a "Taxa de privação material" e o "Rácio S90/S10" são as variáveis independentes mais importantes na explicação da variação da "percentagem da população residente em risco de pobreza ou exclusão social (PPRPES)". A qualidade do ajustamento deste Primeiro Modelo é muito boa, 89% da variabilidade total da variável dependente é explicada pela equação ajustada.

Além do mais, conclui-se que a "carga mediana das despesas associadas à habitação", o "desemprego registado", o "Índice de dependência total" e o "Rácio S80/S20" são as variáveis independentes que têm maior influência na explicação da variação da "Taxa de intensidade da pobreza (TIP)". A qualidade do ajustamento deste Segundo Modelo também é muito boa, o modelo explica 84% das observações da variável dependente.

A segunda parte do estudo, dedicada a estudar a eficácia das respostas sociais em relação à pobreza, permitiu concluir que, no que toca à frequência (número de utentes) das respostas sociais existentes, os Refeitórios ou Cantinas Sociais e os Centros de Alojamento Temporários são as variáveis independentes mais significativas na explicação da variação da percentagem da população residente em risco de pobreza ou exclusão social (PPRPES). A qualidade do ajustamento deste Terceiro Modelo é boa, 77% da variabilidade total da variável PPRPES é explicada pela equação ajustada.

No que toca à capacidade (número de lugares) das mesmas respostas sociais, as variáveis independentes Comunidades de Inserção, Serviços de Atendimento e Acompanhamento Social e Ajuda Alimentar foram as variáveis incluídas no modelo final do Quarto Modelo apresentado, onde a qualidade de ajustamento é muito boa, o modelo explica 93% das observações da percentagem da população residente em risco de pobreza ou exclusão social (PPRPES).

Para além da relevância dos resultados apurados, este trabalho permitiu também detetar a carência de serem realizados mais estudos de análise científica dirigidos a este grupo-alvo "Família e Comunidade", uma vez que, apesar destes se encontrarem em situação de grande vulnerabilidade, sendo dos grupos mais afetados pelo risco de pobreza, o número de estudos encontrados com foco nos mesmos, comparativamente aos restantes grupos, é muito diminuto. Foi este um dos motivos que levou à escolha do objeto deste estudo. Tendo procurado dar um contributo estatístico construtivo, que pretende ter utilidade prática no auxílio da compreensão, antecipação e prevenção deste fenómeno.

Embora as variáveis sejam voláteis e os dados estejam em constante mudança, pode-se concluir que o trabalho obteve os resultados esperados e os modelos podem ser utilizados para melhor compreensão desta realidade e prevenção do fenómeno da pobreza. De salientar que, à semelhança de alguma literatura da especialidade, mencionada na bibliografia, decidiu-se utilizar a Regressão Linear Múltipla e com os modelos estimados foi possível obter valores previstos (pontuais e intervalares) relativamente próximos dos valores reais, disponibilizados pelo INE. Não obstante, poderiam ter sido utilizados modelos econométricos com variável dependente censurada, investigação que poderá vir a ser considerada futuramente.

Bibliografia

Agência Lusa. (2022, 1 novembro). "Cantinas sociais em Lisboa sem procura excecional mantêm número de refeições". *Observador*. Disponível em: https://observador.pt/2022/11/01/cantinas-sociais-em-lisboa-sem-procura-excecional-mantem-numero-de-refeicoes/

Alves, N. 2022. "Um indicador de pobreza multidimensional para Portugal". *Revista de Estudos Económicos*, Vol. VIII, Nº4, páginas 30-54. Disponível em: https://www.bportugal.pt

Barbetta, P. (2010) ESTATISTICA APLICADA ÀS CIÊNCIAS SOCIAIS. (7ª Edição) Editora ufsc

Bioestatística. Análise Descritiva. https://biostatistics-uem.github.io/Bio/descritiva.html

Bruto da Costa, A. (1998) Cadernos democráticos: Exclusões Sociais. Edição Gradiva. Coleção: Fundação Mário Soares.

Carta Social. https://www.cartasocial.pt

Chaudhry, I.S; Malik, S; Hassan, A. (2009, January) "The Impact of Socioeconomic and Demogaphic Variables on Poverty: A Village Study". *The Lahore Journal of Economics*. Bahauddin Zakariya University, Multan, Pakistan. DOI:10.35536/lje.2019.v14.i1.a2

Chein, F. (2019) Introdução aos modelos de regressão linear. Enap- Fundação Escola Nacional de Administração Pública. Coleção: Metodologias de Pesquisa.

Comissão Europeia. https://commission.europa.eu/index_pt

Comissão Europeia. (2021) Fund for European Aid to the Most Deprived https://ec.europa.eu/social/main.jsp?langId=pt&catId=1089

Diogo, F., Farinha Rodrigues, C., Palos, A.C., Pereira, E., Bessa, F., Trevisan, G., Fernandes, L., Silva, O., Perista, P., Amaro, I. "A Pobreza em Portugal: Trajetos e Quotidianos". (2021, abril). *Fundação Francisco Manuel dos Santos*. Disponível em: https://www.ffms.pt/pt-pt/estudos/pobreza-em-portugal-trajetos-e-quotidianos

Dobson, A., Barnett, A. (1990) An introduction to generalized linear models. (4ª Edição) CRC Press

Domingues do Amaral, G., Silva, V.L., Afonso Reis, E. "Análise de Regressão Linear no Pacote R". Relatório técnico série ensino RTE 001/2009. Universidade Federal de Minas Gerais Instituto de Ciências Exatas Departamento de Estatística.

Instituto Nacional de Estatística (INE). https://www.ine.pt

INE. (2020). Inquérito às Condições de Vida e Rendimento - 2022. Lisboa: INE. Disponível em: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE

Gabinete de Estratégia e Estudos. (2022, Setembro) POBREZA E EXCLUSÃO SOCIAL EUROSTAT. https://www.gee.gov.pt/pt/indicadores-diarios/ultimos-indicadores/32334-pobreza-e-exclusao-social-eurostat

Karpinska, L.; Smiech, S.; ´ Gouveia, J.P.; Palma, P. (2021, September) "Mapping Regional Vulnerability to Energy Poverty in Poland". *Sustainability*. Department of Microeconomics, Cracow University of Economics, Cracow, Poland. Department of Statistics, Cracow University of Economics, Cracow, Poland. CENSE—Center for Environmental and Sustainability Research, NOVA School of Science and Technology, NOVA University Lisbon. DOI: https://doi.org/10.3390/ su131910694

Ministério do Trabalho, Solidariedade e Segurança Social. (2021) Carta Social – *Rede de Serviços e Equipamentos, Relatório 2021*

Mendes, M.F., Pinto, J.E., Rebelo dos Santos, J. "A comparative analysis of poverty among the poorest countries in the European Union, a specific glance on Portugal". (2005, September) Trabalho apresentado no *XI Meeting of APDR*. Faro, Portugal. Disponível em: http://hdl.handle.net/10400.26/20582

Murteira, B., Ribeiro, C., Andrade e Silva, J., Pimenta, C., Pimenta, F. (2015) Introdução à Estatística. (4ª Edição) Escolar Editora

Observatório das Desigualdades. (junho, 2018) *"Rácio S90/S10 (desigualdade de rendimento)"*. Disponível em: https://observatoriodasdesigualdade.wordpress.com/

Pereira, D. (2022) *Regressão Linear e Multilinear* [PowerPoint slides]

Porfírio, J. (2021, 12 abril). "Um quinto da população portuguesa é pobre. Um em cada três pobres tem emprego estável". *Observador*. Disponível em: https://observador.pt/2021/04/12/a-pobreza-em-portugal-quase-20-das-pessoas-sao-pobres-e-um-em-cada-tres-pobres-tem-emprego-estavel/

Rodrigues, C.F. (coord.), Figueiras, R., Junqueira, V. (2016), *Desigualdade do Rendimento e Pobreza em Portugal*, Lisboa, FFMS.

Sinaga, M. "Analysis of Effect of GRDP (Gross Regional Domestic Product) Per Capita, Inequality Distribution Income, Unemployment and Human Development Index on Poverty" (2018) Faculty of Economic and Business, Universitas Sumatera Utara, Indonesia. DOI: https://doi.org/10.33258/birci.v3i3.1177