INSTITUTO UNIVERSITÁRIO DE LISBOA

Previsão da percentagem da despesa total direcionada para alimentação num agregado familiar de Espanha (por mês)

Licenciatura em Ciência de Dados (2º ano)

Introdução a Modelos Dinâmicos

Docente: Maria da Conceição Figueiredo

Grupo 4:

Gonçalo Mealha 123391

Jingyu Huang 123432

Ricardo Valério 112255

Índice

Introdução

No âmbito da unidade curricular de Introdução a Modelos Dinâmicos, foi proposto o desenvolvimento de um modelo de regressão múltipla que permita estimar a percentagem da despesa que é gasta na alimentação por família, com base nas variáveis contidas na base de dados dada pelos docentes.

Com o objetivo de melhorar o desempenho e a organização do trabalho, o grupo irá seguir a metodologia de CRISP-DM (*Cross Industry Standard Process for Data Mining*), conforme ilustrada na figura abaixo. Como representado no gráfico, a etapa de modelação é precedida por um processo iterativo de preparação dos dados. Esta última é uma fase crítica que envolve a seleção, limpeza e transformação das variáveis, de modo a garantir que os dados estejam na melhor forma possível para a construção do modelo. A execução adequada desta etapa é fundamental para assegurar a qualidade e a fiabilidade dos resultados obtidos na fase de modelação.

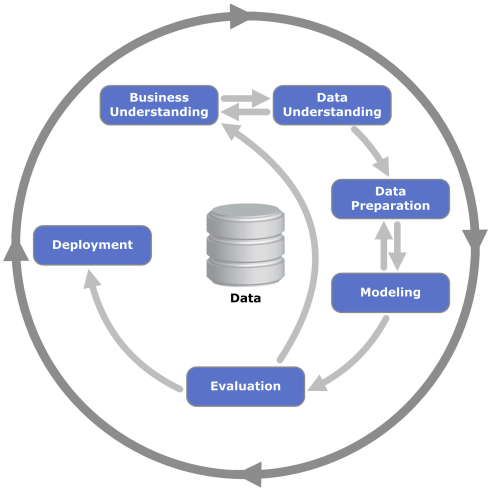

Fig. 1 - CRISP - DM

Business Understanding

O objetivo deste trabalho é desenvolver um modelo de regressão que permita estimar a percentagem da despesa total de um agregado familiar espanhol direcionada para alimentação.

A previsão dessa percentagem é extremamente importante para compreender como as famílias em Espanha distribuem o seu orçamento e como fatores demográficos, sociais e geográficos influenciam esse comportamento de consumo.

A partir do modelo de regressão desenvolvido, será possível identificar quais as variáveis que apresentam maior impacto nos gastos na alimentação, e assim, extrair *insights* valiosos sobre como os diferentes fatores influenciam os padrões de consumo das famílias espanholas.

Data Understanding

O conjunto de dados "BudgetFood[1]" é resultado de um estudo transversal (*cross-section*) conduzido na Espanha em 1980, com foco nas despesas domésticas, especificamente a percentagem do gasto total destinado à alimentação. Este conjunto de dados inclui 23.972 observações, onde cada observação representa uma família. Neste caso, foi atribuída ao nosso grupo uma subamostra com 2501 observações, composta por 7 variáveis distintas, sendo elas:

Atributos	Tipo	Conteúdo
X	int	ID das observações (famílias)
wfood	num	Percentagem da despesa total que uma família gasta na alimentação
totexp	num	Despesas totais do agregado familiar
age	num	Idade da pessoa de referência do agregado familiar
size	int	Número de agregado familiar
town	int	Tamanho da cidade onde a família tem o domicílio (5 categorias: 1 - cidade pequena, … , 5 - cidade grande)
sex	Factor	Sexo da pessoa da referência do agregado familiar (man, woman)

Com essas variáveis, a base de dados oferece uma visão abrangente das famílias de Espanha, incluindo informações sobre seus gastos, composição e características demográficas.

Limpeza e Transformação dos Dados

Antes de importar os dados para o R, o grupo analisou previamente as variáveis presentes no conjunto. Nesse contexto, no processo de importação, a coluna designada por *X*, correspondente ao ID das observações, foi imediatamente

[1] https://rdrr.io/cran/Ecdat/man/BudgetFood.html

descartada, visto que não apresenta qualquer relevância para as análises subsequentes.

```
# Importar a base de dados do ficheiro CSV
budget_food_data <- read.csv(
                        file = "BudgetFood_Grupo4.csv",
                        sep = ";",
                        stringsAsFactors = TRUE
                    )[ , -1] # ignorar a 1ª coluna (índice de observações)
```

Fig. 2 - Importação da base de dados excluindo a coluna X

É importante destacar que a presença de valores nulos ou duplicados pode comprometer a qualidade e a integridade dos dados, dificultando as análises subsequentes e, por consequência, o modelo desenvolvido.

Dessa forma, foi desenvolvido um código para verificar a presença de valores omissos (*NA)* na base de dados, e, felizmente, constatou-se que não existem valores ausentes.

```
# Verificar a existência dos valores nulos
colSums(is.na(budget_food_data))
```

```
wfood:    0 totexp:    0 age:    0 size:    0 town:    0 sex:    0
```

Fig. 3 - Verificação dos NA's

Quanto aos dados duplicados, foram detectadas 6 observações duplicadas, as quais foram removidas para garantir a integridade do conjunto de dados e assegurar a qualidade das análises subsequentes.

	wfood	totexp	age	size	town	sex
	<dbl>	<int>	<int>	<int>	<int>	<fct>
1609	0.09017669	4384814	52	4	4	man
1610	0.14984846	3278272	45	6	4	man
1611	0.19597145	904295	41	2	4	woman
1612	0.12740736	2051310	50	4	4	man
1613	0.06305279	2769362	33	4	4	man
1614	0.21822980	1132311	45	5	4	man
1632	0.09017669	4384814	52	4	4	man
1633	0.14984846	3278272	45	6	4	man
1634	0.19597145	904295	41	2	4	woman
1635	0.12740736	2051310	50	4	4	man
1636	0.06305279	2769362	33	4	4	man
1637	0.21822980	1132311	45	5	4	man

Fig. 4 - Registos duplicados

Na base de dados, está presente uma variável categórica - ***sex****,* que assume os valores *man* e *woman.* Para possibilitar a integração dessa variável no modelo, foi

transformada numa variável *dummy*, onde o valor 1 corresponde a *man* e o valor 0 corresponde a *woman*.

Como a variável ***totexp*** está representada em Peseta Espanhola, foi necessário convertê-la para Euro. Sabendo que 1000 Pesetas equivalem a 6 Euros, a conversão foi realizada multiplicando os valores da coluna *totexp* por 6 e, em seguida, dividir o resultado por 1000.

```
# Transformar a coluna "totexp" (em Peseta Espanhola) para Euros(€)
# sabendo que 1000 pesetas = 6 euros
budget_food_data$totexp <- budget_food_data$totexp * 6 / 1000
```

Fig. 6 - Comando para conversão da coluna *totexp*

	wfood	totexp	age	size	town	sex
	<dbl>	<dbl>	<int>	<int>	<int>	<dbl>
1	0.2978185	5568.090	30	6	2	1
2	0.2433597	8415.396	56	4	2	1
3	0.6531605	844.056	44	1	2	1

Fig. 7 - Dados após a conversão

Feitas estas pequenas alterações, o grupo pôde começar a explorar os dados.

Análise Descritiva

Boxplots, Histogramas e Barplots

Começámos por extrair as medidas descritivas das variáveis de forma a adquirir alguma sensibilidade à base de dados.

```
     wfood             totexp             age              size          town          sex        
 Min.   :0.01504   Min.   :  121.7   Min.   :18.00   Min.   : 1.000   4:870   Min.   :0.0000  
 1st Qu.:0.26063   1st Qu.: 2638.9   1st Qu.:39.00   1st Qu.: 2.000   1:441   1st Qu.:1.0000  
 Median :0.36083   Median : 4372.4   Median :51.00   Median : 4.000   2:599   Median :1.0000  
 Mean   :0.37822   Mean   : 5167.8   Mean   :51.28   Mean   : 3.706   3:580   Mean   :0.8695  
 3rd Qu.:0.47952   3rd Qu.: 6556.4   3rd Qu.:62.00   3rd Qu.: 5.000           3rd Qu.:1.0000  
 Max.   :0.95696   Max.   :68385.3   Max.   :97.00   Max.   :17.000           Max.   :1.0000  
```

Fig.8 - summary() da base de dados

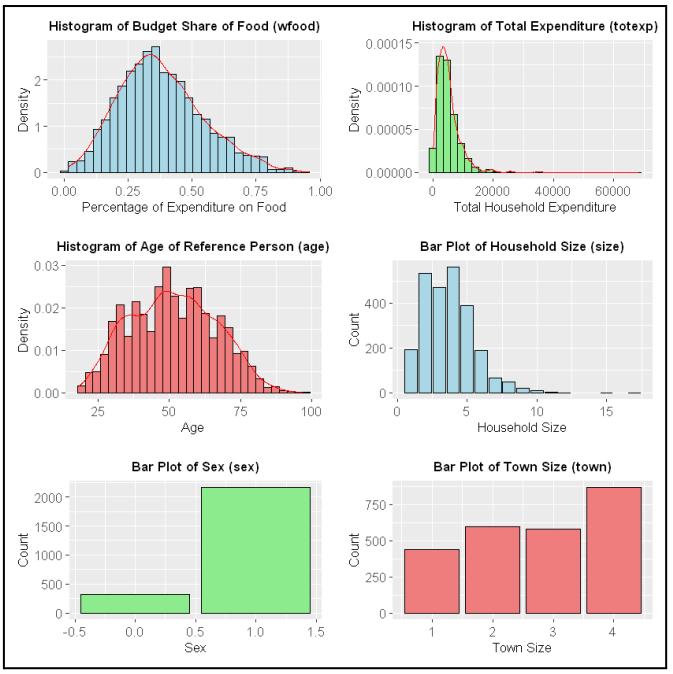

Fig.9 - Histogramas e barplots para as variáveis da base de dados

Foi possível retirar algumas conclusões, nomeadamente:

wfood: Corresponde à percentagem da despesa total que uma família gasta para alimentação. Varia de 0.0 (nenhuma despesa com alimentação) a 0.957, com a média de 0.3769.

O histograma apresenta uma distribuição simétrica, indicando que a maioria das famílias destina uma proporção equilibrada da sua renda às despesas com alimentação. Isto é, o gasto com comida não está excessivamente concentrado em valores muito baixos ou muito altos, refletindo um padrão consistente entre os lares.

totexp: Os gastos totais variam de 121.7 a 68385.3 (euros), sendo a média de 5183. É, portanto, uma variável de grande amplitude que muito provavelmente contém *outliers.*

O histograma é assimétrico à direita, indicando que a maioria das famílias tende a ter despesas totais mais baixas, enquanto um número menor de famílias gasta significativamente mais. Economicamente, isso sugere que, enquanto a maioria das famílias opera dentro de um orçamento mais modesto, existe um subconjunto de famílias mais ricas com uma capacidade de gasto muito maior.

age: A amostra inclui apenas indivíduos de idade adulta, com idades entre os 18 e os 97 anos e uma média de 51 anos. O histograma é aproximadamente simétrico, indicando que a distribuição etária no conjunto de dados está bem distribuída em torno da média, sem uma tendência significativa para a esquerda ou direita. Isso sugere uma representação equilibrada de diferentes faixas etárias na amostra, incluindo jovens adultos, adultos de meia-idade e idosos.
Como a idade influencia os hábitos de consumo - com diferenças entre pessoas mais jovens e mais velhas - essa distribuição equilibrada possibilita uma análise mais precisa do impacto da idade nas decisões económicas, reduzindo possíveis distorções que poderiam surgir caso a amostra fosse enviesada para um grupo etário específico.

size: O número de elementos do agregado familiar varia de 1 a 17 membros, possuindo, em média, aproximadamente 4 pessoas. A distribuição do tamanho do agregado familiar é assimétrica à direita, com a maioria das observações concentradas em valores baixos (entre 1 e 6 membros).

town: A dimensão das cidades varia entre 1 e 4, não havendo nenhum registo de uma “cidade grande”(5).
O gráfico de barras mostra que a maioria das famílias vive em cidades de categoria 4, seguidas pelas de categoria 2 e 3, sugerindo uma concentração populacional em áreas urbanas tendencialmente maiores.

sex: Como a média da variável está bastante próxima de 1, a maioria da amostra é composta por homens (2173 indivíduos), enquanto as mulheres representam apenas 328 indivíduos.
O gráfico de barras, tanto na nossa subamostra quanto em todo o *dataset*, diz-nos que o género predominante da pessoa de referência no domicílio é identificado como homem, indicando que os homens são maioritariamente considerados os “chefes de família”. Esse padrão pode refletir papéis tradicionais de caráter histórico dentro das estruturas familiares ao longo dos tempos, nos quais os homens tendencialmente costumam assumir responsabilidades de tomada de decisões dentro do lar. E portanto, de certa forma, isso pode talvez impactar a forma como os recursos são distribuídos dentro do domicílio, incluindo

a percentagem do orçamento destinada à alimentação (*wfood*) e outros gastos domésticos.

Para verificar a existência de valores extremos (*outliers*) foi construído para cada variável (à exceção da variável *sex* que só apresenta 2 categorias de resposta) um boxplot:

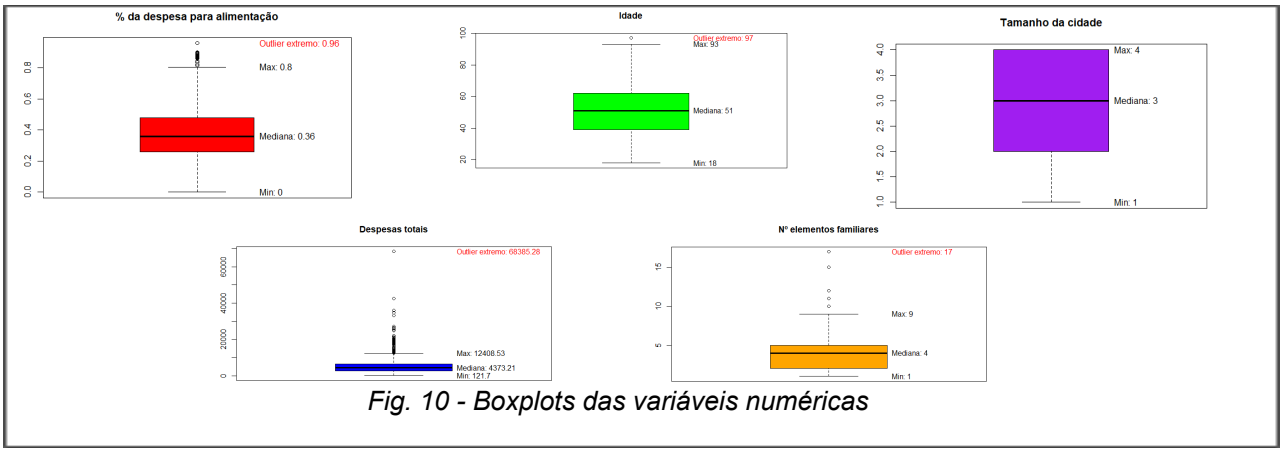

Fig. 10 - Boxplots das variáveis numéricas

wfood - A presença de alguns valores extremos na extremidade superior, mostra que algumas famílias dedicam uma percentagem significativamente maior do seu orçamento para os alimentos. Isso pode representar famílias de baixa renda onde a alimentação ocupa uma grande proporção das suas despesas ou agregados familiares de grande dimensão.

totexp - Existem vários *outliers* superiores, que representam famílias com elevados rendimentos e com despesas significativamente maiores. Destaca-se uma observação com 68385.3 euros gastos.

age - Não existem valores extremos, muito devido à distribuição normal da variável. Essa distribuição sugere uma distribuição demográfica onde nenhuma faixa etária específica domina o conjunto de dados.

size - Existem 17 famílias de grande dimensão cujo agregado familiar é superior a 9 pessoas.

1	2	3	4	5	6	7	8	9	10	11	12	15	17
196	535	470	562	391	188	66	49	21	11	3	1	1	1

Fig. 11 - Tabela de Frequências para a variável size

town - Devido ao seu reduzido número de categorias, e por consistir numa variável com dados do tipo escala de *Likert* - a variável *town* não apresenta quaisquer *outliers*.

Ainda na variável ***wfood***, deparámo-nos com uma situação irregular onde haviam famílias que não gastavam nenhuma parte do seu dinheiro em alimentação (*wfood* = 0). Ora, esta situação é estranha pois todas as pessoas necessitam de comida para sobreviver. Vamos analisar estes casos:

	wfood	totexp	age	size	town	sex
947	0	454.560	72	1	2	0
998	0	798.996	78	1	2	1
1455	0	3256.440	26	3	2	1
1913	0	1806.144	55	2	4	0
2018	0	1355.604	49	1	4	1

Fig.12 - Famílias sem despesas em alimentação

Sendo um número ínfimo de casos neste cenário, à partida a principal razão para tal ter acontecido seria um erro de digitação dos dados. No entanto, podem haver outras razões. No caso das duas pessoas idosas (observação 947 e 998), poderiam ser duas pessoas já reformadas que se encontram dependentes de outrem e embora tenham sido registadas como a pessoa de referência do agregado familiar, são, portanto, sustentadas por outras pessoas. Nos restantes casos poderiam ser pessoas necessitadas que recebiam refeições gratuitas e doações, porém o valor elevado de despesas totais refuta essa hipótese. Seja qual for o cenário, estas são situações atípicas e portanto o grupo decidiu tratar estas observações como *outliers* e consequentemente removê-las.

Matriz de Dispersão e Matriz de Correlações

Antes de avançar para o cálculo das correlações é importante perceber que tipo de relação existe entre as variáveis. Para isso, recorremos à criação de gráficos de dispersão para cada par de variáveis.

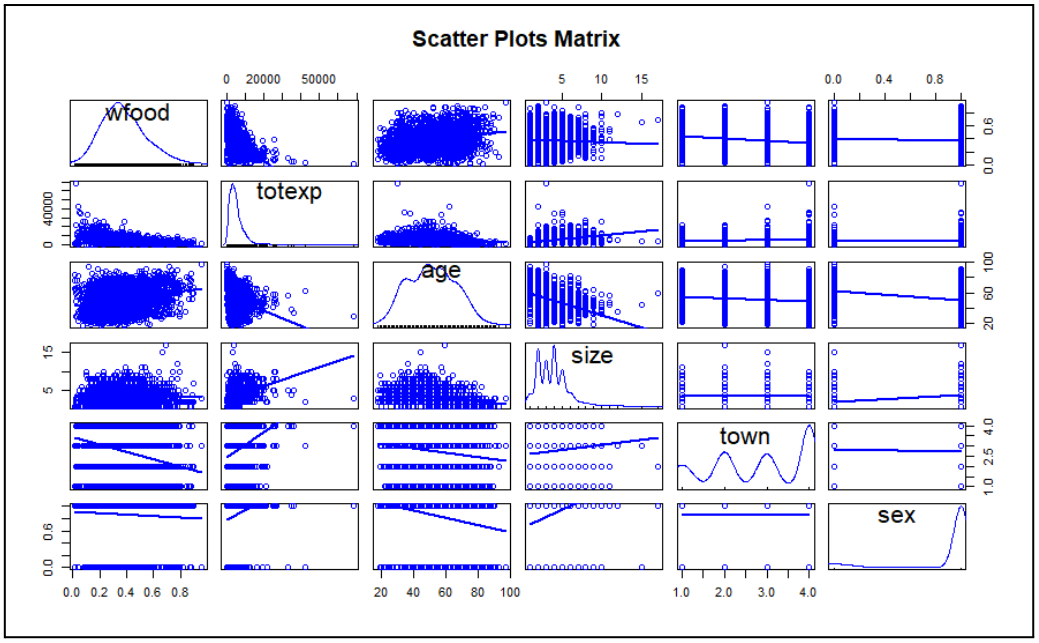

Fig. 13 - Matriz de dispersão

A relação mais evidente ocorre entre as despesas totais (*totexp*) e a percentagem gasta em alimentação (*wfood*), mostrando que famílias mais abastadas gastam proporcionalmente menos com comida. Contudo, a relação entre as variáveis parece ter contornos não lineares, ou seja, à medida que as despesas totais aumentam as diferenças entre as percentagens direcionadas aos gastos com a alimentação vão sendo menos acentuadas.

Parece haver uma leve tendência positiva entre a idade (*age*) e os gastos na alimentação (*wfood*), no entanto existe uma alta dispersão nos dados.

Não parece haver uma relação clara entre a dimensão do agregado familiar (*size*) e os gastos na alimentação (*wfood*) mas há alguns valores dispersos, possivelmente sugerindo que o tamanho da família pode ter um efeito não linear.

Para complementar a análise da relação entre as variáveis foi construída a matriz de correlações.

	wfood	totexp	age	size	town	sex
wfood	1.00000000	-0.5082536	0.2573123	-0.04663193	-0.26754086	-0.06302258
totexp	-0.50825357	1.0000000	-0.2623574	0.37126596	0.27174155	0.19201889
age	0.25731229	-0.2623574	1.0000000	-0.38470869	-0.13609020	-0.27012098
size	-0.04663193	0.3712660	-0.3847087	1.00000000	0.08750351	0.32155099
town	-0.26754086	0.2717415	-0.1360902	0.08750351	1.00000000	-0.01993470
sex	-0.06302258	0.1920189	-0.2701210	0.32155099	-0.01993470	1.00000000

Fig. 14 - Matriz de correlações

As correlações entre a variável alvo e as variáveis independentes são, geralmente, fracas. A relação mais forte ocorre entre ***totexp*** e ***wfood*** que, embora seja uma correlação negativa moderada, ela sugere que, à medida que a despesa total aumenta, a proporção da despesa gasta na alimentação tende a diminuir tal como identificado no *scatter plot*. Intrigados por esta relação, pesquisámos melhor sobre as possíveis causas e descobrimos que existem fenómenos económicos estudados que explicam esta associação. Nomeadamente, a *Lei de Engel*[2], que afirma que, nas famílias com maior poder de compra, a fatia percentual destinada à alimentação costuma ser menor, pois elas também tendem a investir em outras áreas (lazer, educação, saúde, etc.).

Continuando com a nossa análise, verificamos que o tamanho da cidade em que o inquirido habita (*town*) está negativamente correlacionado com os gastos na alimentação (*wfood*). Isso pode estar associado ao facto de que o elevado custo de vida nas cidades grandes (como indicado pela correlação positiva entre *town* e as despesas totais do agregado familiar (*totexp*), faz com que as despesas sejam direcionadas a outro tipo de necessidades como por exemplo a habitação ou os serviços de transporte, fazendo com que seja desembolsada uma menor fatia das despesas para a alimentação.

[2] https://en.wikipedia.org/wiki/Engel%27s_law

Segue-se a idade (*age*), que é positivamente correlacionada com a *wfood* (0,25), sugerindo que pessoas mais velhas tendem a direcionar uma percentagem maior do seu orçamento para os alimentos.

Tal como indicado no *scatter plot,* a correlação entre *wfood* e o tamanho do agregado familiar (*size*) é muito fraca e praticamente inexistente (-0,04). O mesmo acontece para o sexo do inquirido (*sex*) que apresenta uma correlação de apenas -0.06, o que pode se dever à baixa representatividade do sexo feminino na amostra.

Não existem correlações significativas entre as variáveis independentes, sendo a mais elevada de apenas -0,38, ou seja, à partida não existirá multicolinearidade entre as variáveis preditoras.

Modelos

Para cada modelo desenvolvido, serão analisados os seguintes parâmetros: o valor do R^2, utilizado como a métrica para avaliar a eficácia do modelo; o valor de *AIC* (*Akaike*) e por último, a verificação dos pressupostos dos resíduos, que são os seguintes:

1.A média da distribuição de probabilidades de ε é zero:

$$E(\varepsilon) = 0, \forall x \rightarrow E(y) = \beta_0 + \beta_1 x$$

2. Homocedasticidade (variância constante):

$$Var(\varepsilon_i) = \sigma_\epsilon^2 = \sigma^2, \forall x$$

3. Não existe autocorrelação entre ε_i e ε_j :

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \ (i \neq j)$$

4. A distribuição de probabilidades de ε é normal:

$$\varepsilon \sim N\ (0, \sigma^2)$$

5. A variável explicativa é independente dos resíduos:

$$Cov\ (\varepsilon_i, x_j) = 0$$

A análise dos resíduos permite avaliar se o modelo assumido é adequado (ou seja, se o modelo se ajusta da melhor forma aos dados em consideração).

Modelo 1

O grupo começou pelo desenvolvimento de um modelo de regressão linear múltipla, denominado *modelo1*, que inclui todas as variáveis da base de dados sem quaisquer transformações.

```
Call:
lm(formula = wfood ~ ., data = budget_food_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47722 -0.08708 -0.01473  0.07892  1.05604

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.573e-01  1.774e-02  20.139   <2e-16 ***
totexp      -2.102e-05  7.463e-07 -28.164   <2e-16 ***
age          2.108e-03  1.994e-04  10.568   <2e-16 ***
size         2.053e-02  1.733e-03  11.846   <2e-16 ***
town        -2.163e-02  2.487e-03  -8.696   <2e-16 ***
sex          5.765e-03  8.561e-03   0.673    0.501
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1336 on 2484 degrees of freedom
Multiple R-squared:  0.3375,    Adjusted R-squared:  0.3362
F-statistic: 253.1 on 5 and 2484 DF,  p-value: < 2.2e-16
```

Fig.15 - Sumário do Modelo 1

Os resultados do modelo1 indicam que cerca de 33,75% da variabilidade na percentagem da despesa direccionada para alimentação é explicada pela variação das variáveis independentes (R^2= 0.3375 -> coeficiente de determinação). Este valor sugere que o modelo apresenta uma capacidade explicativa fraca.

O teste F avalia a significância global do modelo, determinando se no conjunto, as variáveis explicativas (ou seja, todas excepto o termo independente β_0) são ou não relevantes para a explicação da variável dependente. Neste caso, como o *p-value* do teste F é menor do que 2.2e-16, rejeita-se a hipótese nula de que todas as variáveis não têm efeito na variável dependente e portanto, existe pelo menos uma variável

independente (β_j) que tem um impacto não nulo na percentagem da despesa atribuída para a alimentação.

Através da análise dos coeficientes de cada uma das variáveis independentes, verificou-se que, à exceção da variável *sex,* todas as restantes são relevantes para o modelo. A variável *sex* apresenta um *p-value* bastante alto de 0.501, indicando que não é estatisticamente significativa a 10%. Para este modelo, o valor de *AIC* registado foi de -2948.77.

Quanto aos pressupostos, o único que se verifica, na íntegra, isto é, que é confirmado pelo teste é o da média (nula) dos resíduos. Apesar de o pressuposto da normalidade dos resíduos não ter sido verificado pelo teste de *Jarque-Bera,* como a amostra é suficientemente grande, podemos invocar o Teorema do Limite Central (TLC) e assumir que os resíduos são normalmente distribuídos. Esta assunção será feita para todos modelos e portanto todos eles verificarão o quarto pressuposto independentemente do resultado do teste de *Jarque-Bera*.

Modelo 2

Apesar da variável *sex* não ser estatisticamente significativa no *modelo1*, decidimos mantê-la para a construção do *modelo2,* pois a sua remoção não alterou em nada o valor do R^2 nem o *AIC*. Assim, para a construção do *modelo2* o grupo decidiu transformar a variável town numa variável *dummy* através da função *as.factor()*, onde a categoria 4, por ser a mais frequente, foi escolhida como a categoria de referência. Além disso, devido à grande amplitude da variável *totexp*, a mesma foi logaritmizada na tentativa de reduzir a sua dispersão.

```
Call:
lm(formula = wfood ~ log(totexp) + age + size + sex + as.factor(town),
    data = budget_food_data)
```

```
Residual standard error: 0.1281 on 2482 degrees of freedom
Multiple R-squared:  0.3917,    Adjusted R-squared:   0.39
F-statistic: 228.4 on 7 and 2482 DF,  p-value: < 2.2e-16
```

Fig. 16 - Modelo 2 e resultados parciais

Houve uma melhoria na percentagem de variância da variável alvo explicada pelo modelo que passou a 39,17%. Agora, todas as variáveis preditoras são estatisticamente significativas a pelo menos 1% (0.01). O valor de *AIC* também sofreu uma ligeira melhoria, tendo diminuído para -3157.281.
Os únicos pressupostos a serem verificados continuam a ser o da média dos resíduos e da sua normalidade.

Modelo 3

Neste modelo, resolvemos adicionar 2 interações entre variáveis com o objetivo de poder captar novas estruturas de dados.

log(totexp):size - Esta interação ajuda a captar como o efeito da despesa total varia conforme o tamanho da família.
log(totexp):town - Esta interação é importante para captar diferenças regionais no que toca às despesas.

```
Call:
lm(formula = wfood ~ log(totexp) + sex + age + log(totexp):size + 
    log(totexp):as.factor(town), data = budget_food_data)
```

```
Residual standard error: 0.1286 on 2482 degrees of freedom
Multiple R-squared:  0.3873,    Adjusted R-squared:  0.3856 
F-statistic: 224.2 on 7 and 2482 DF,  p-value: < 2.2e-16
```

Fig. 17 - Modelo 3 e resultados parciais

0 R^2 apresentou uma ligeira diminuição em relação ao modelo anterior, bem como o valor de *AIC* piorou face ao anterior, passando a ser de -3139.257. Apesar das novas alterações, os pressupostos da variância (constante) dos erros e da independência entre eles continuam por verificar. Neste modelo, bem como nos anteriores, existe a presença de heterocedasticidade nos dados, ou seja, a variação dos resíduos não é constante ao longo das observações. Esse problema é bastante evidente, através da análise do gráfico dos resíduos *vs* valores ajustados que apresenta uma forma de funil.

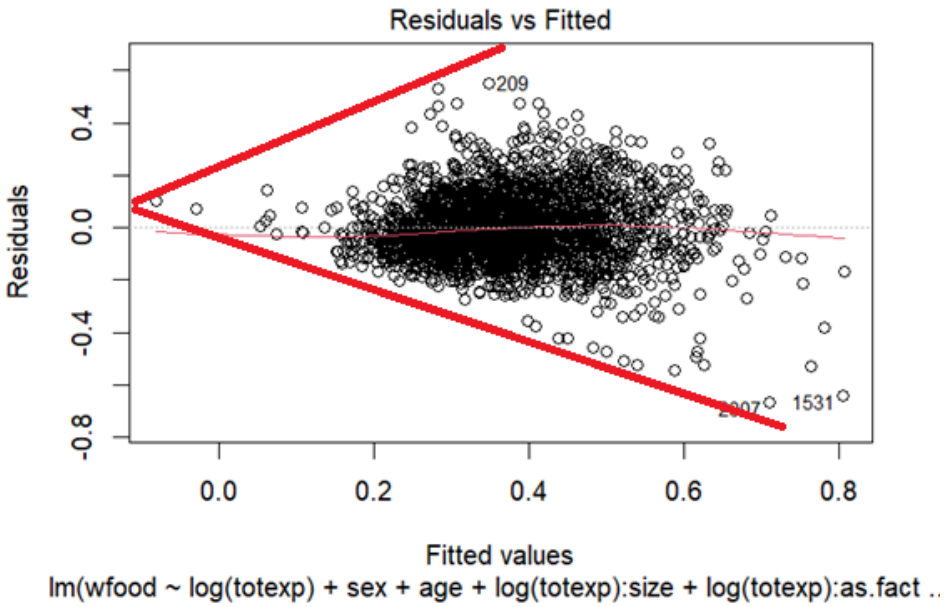

Fig. 18 - Heterocedasticidade no Modelo 3 (forma de funil)

Modelo 4

Para lidar com o problema da heterocedasticidade, optámos por ajustar um modelo de regressão ponderado (*WLS - Weighted Least Squares*), onde os pesos são estimados a partir de um modelo de regressão robusto (*RLM - Robust Linear Model*). O objetivo é estimar pesos para cada observação, de forma que as observações com resíduos maiores tenham menor impacto no modelo final, ao contrário do Método dos Mínimos Quadrados (OLS) em que todas as observações têm o mesmo impacto no modelo.

Foi usado um modelo auxiliar robusto (modelo4_auxiliar) para calcular os pesos usados no modelo de regressão (modelo4). Ambos os modelos tiveram a mesma fórmula, isto é, as mesmas variáveis preditoras.

```
Call:
lm(formula = wfood ~ log(totexp) + sex + age + log(totexp):size +
    log(totexp):as.factor(town), data = budget_food_data, weights = weights)
```

```
Residual standard error: 0.1081 on 2482 degrees of freedom
Multiple R-squared:  0.4793,    Adjusted R-squared:  0.4778
F-statistic: 326.3 on 7 and 2482 DF,  p-value: < 2.2e-16
```

Fig. 19 - Modelo 4

O valor do *R-squared* aumentou significativamente, indicando que o modelo agora explica aproximadamente 48% da variabilidade da variável dependente. O valor do *AIC* ficou ainda mais negativo (-3821.543).
No entanto, apesar das transformações realizadas com o objetivo de minimizar a heterocedasticidade, no teste de *Breusch-Godfrey* continuámos a rejeitar a hipótese nula de que o modelo é homocedástico e portanto a heterocedasticidade persiste nos dados. Os únicos pressupostos que conseguimos verificar até ao momento foram o da média dos resíduos e da normalidade dos mesmos.

Modelo 5

No modelo5 o grupo decidiu explorar outra abordagem. A variável dependente foi transformada, multiplicando a percentagem da despesa total que é gasta em comida pelas despesas totais (*wfood * totexp*). Assim, o modelo passa a prever o valor total gasto em alimentação, em vez da proporção desse gasto em relação ao orçamento total. É preciso ter em conta que esta transformação altera a interpretação dos coeficientes, pois agora eles refletem o impacto das variáveis explicativas sobre o valor absoluto gasto em alimentação, em vez da fração do orçamento destinada a esse fim. Dado que a variável *totexp* possui uma grande amplitude de valores, aplicámos a transformação logarítmica à nova variável dependente *(log(wfood*totexp))*.
Como identificado anteriormente no gráfico de dispersão entre *wfood* e *totexp*, a relação entre estas variáveis não era totalmente linear e existia uma curvatura nos dados. Assim, decidimos adicionar um novo termo *(totexp*^2) de segundo grau de forma a capturar essa tendência. Além disso, utilizámos pesos de *1/sqrt(totexp)*, que reduzem a influência de observações com despesas totais muito elevadas e dão maior peso às de menor despesa. Essa ponderação procura estabilizar a variância dos resíduos e melhorar a robustez do modelo.

```
Call:
lm(formula = log(wfood * totexp) ~ log(totexp) + I(totexp^2) + 
    sex + age + log(totexp):size + log(totexp):as.factor(town), 
    data = budget_food_data, weights = 1/sqrt(totexp))
```

```
Residual standard error: 0.05738 on 2481 degrees of freedom
Multiple R-squared:  0.6847,    Adjusted R-squared:  0.6837
F-statistic: 673.6 on 8 and 2481 DF,  p-value: < 2.2e-16
```

Fig. 20 - Modelo 5

O modelo5 apresenta um *R-Squared* de 0.6847, sendo o mais elevado até ao momento. Por outro lado, o valor do *AIC* passou a positivo, devido, provavelmente, a ter logaritmizado a variável dependente.

```
        studentized Breusch-Pagan test

data:  modelo5
BP = 5.8783, df = 8, p-value = 0.6609
```

Fig. 21 - Teste de *Breusch-Pagan* para o Modelo5

Como o *p-value* do teste de *Breusch-Pagan* é superior a 0.05 (e a 0.1), não rejeitamos a hipótese nula de que a variância dos resíduos é constante e, portanto, o pressuposto é verificado, demonstrando que a utilização dos pesos foi adequada. Desta forma, o único pressuposto que falta verificar é o da independência entre os resíduos.

Modelo 6

No *modelo6*, o grupo adotou uma nova abordagem ao redefinir a variável dependente como a despesa alimentar *per capita*, calculada através da razão *(wfood/size).* Dessa forma, o modelo passa a explicar a proporção do rendimento gasto em alimentação por cada membro do agregado familiar, em vez do valor total ou da percentagem do orçamento destinada a esse fim. A inclusão da variável *size* como fator de ajuste permite captar diferenças no consumo alimentar entre famílias de diferentes dimensões. Ademais, foi adicionada uma nova interação entre a variável *size* e *age* para captar como o efeito da idade do chefe de família varia conforme o tamanho do agregado familiar (*size:age*).

Como os pesos *(1/sqrt(totexp))* utilizados no *modelo5* reduziram a heterocedasticidade dos resíduos, decidimos aplicá-los também neste modelo.

Com estas modificações, procurámos validar o pressuposto que nos falta: o da ausência de autocorrelação residual.

```
Call:
lm(formula = ((wfood)/size) ~ totexp + age + sex + as.factor(town) +
    size:age + (totexp:size) + (totexp:as.factor(town)), data = budget_food_data,
    weights = 1/sqrt(totexp))

Weighted Residuals:
      Min        1Q    Median        3Q       Max
-0.077261 -0.004410 -0.000504  0.003086  0.094738

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.040e-01  1.033e-02  19.742  < 2e-16 ***
totexp                  -3.009e-05  1.540e-06 -19.537  < 2e-16 ***
age                      4.189e-03  1.354e-04  30.942  < 2e-16 ***
sex                     -5.944e-02  5.333e-03 -11.145  < 2e-16 ***
as.factor(town)1         4.144e-02  8.344e-03   4.966 7.28e-07 ***
as.factor(town)2         3.118e-03  7.727e-03   0.404   0.6866
as.factor(town)3        -4.995e-03  8.267e-03  -0.604   0.5458
age:size                -1.034e-03  3.285e-05 -31.483  < 2e-16 ***
totexp:size              4.876e-06  2.745e-07  17.765  < 2e-16 ***
totexp:as.factor(town)1 -4.582e-06  1.855e-06  -2.469   0.0136 *
totexp:as.factor(town)2  9.430e-07  1.580e-06   0.597   0.5507
totexp:as.factor(town)3  1.722e-06  1.443e-06   1.193   0.2329
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01153 on 2478 degrees of freedom
Multiple R-squared:  0.649,     Adjusted R-squared:  0.6474
F-statistic: 416.5 on 11 and 2478 DF,  p-value: < 2.2e-16
```

Fig. 22 - Sumário do Modelo 6

Apesar de o valor do R^2 ser inferior ao do *modelo5*, o modelo apresenta um ajuste onde cerca de 65% da variabilidade da despesa alimentar *per capita* é explicada pelo modelo. Por outro lado, o valor do *AIC* foi o melhor registado até ao momento (-4813.159).

A variância residual continua a ser constante, o que indica que a escolha dos pesos utilizados foi apropriada.

```
        Breusch-Godfrey test for serial correlation of order up to 1

data:  modelo6
LM test = 3.0672, df = 1, p-value = 0.07989
```

Fig. 23 - Teste de *Breusch-Godfrey* para o Modelo 6

Em relação ao pressuposto da autocorrelação entre os resíduos, como o *p-value* do teste de *Breusch-Godfrey* é superior a 0.05, não rejeitamos a hipótese nula de que

os resíduos sejam independentes e, portanto, é o primeiro modelo que verificou o terceiro pressuposto.

Discussão dos modelos com base no R^2, AIC e pressupostos

Como forma de comparação, foi construída uma tabela com todos os modelos e os respectivos *R-squared*, valores de *AIC* e o cumprimento dos pressupostos.

Modelo	R^2	AIC	Pressupostos			
			Média nula	Homocedasticidade	Independência	Normalidade
Modelo1	0.3375	-2349	SIM	NÃO	NÃO	SIM
Modelo2	0.3917	-3157	SIM	NÃO	NÃO	SIM
Modelo3	0.3873	-3139	SIM	NÃO	NÃO	SIM
Modelo4	0.4793	-3822	SIM	NÃO	NÃO	SIM
Modelo5	0.6847	3175	SIM	SIM	NÃO	SIM
Modelo6	0.6490	-4815	SIM	SIM	SIM	SIM

Os primeiros quatro modelos (*modelo1* a *modelo4*) apresentaram valores de *R²* relativamente baixos, variando de 0.3375 a 0.4793, indicando que explicam apenas uma parte da variabilidade de *wfood*. Entre eles, o modelo4 foi o que obteve o melhor ajuste (*R²* = 0.4793) e o menor *AIC* (-3822), indicando ser o mais eficiente destes modelos. No entanto, todos estes modelos sofrem de problemas de heterocedasticidade e presença de autocorrelação dos resíduos, o que pode comprometer sua validade.

Nos modelos 5 e 6 foram aplicadas transformações nas variáveis dependentes e imputados novos pesos. O *modelo5*, ao utilizar a transformação logarítmica na variável dependente, obteve o maior R² (0.6847). No entanto, o valor de *AIC* (3175) foi muito alto, o que pode indicar que a (maior) complexidade do modelo pode não

ser vantajosa. Além disso, a independência dos resíduos não foi plenamente atendida.

Já o *modelo6*, que ajusta o consumo per capita, apresentou um R^2 competitivo (0.6490), o menor *AIC* de todos os modelos (-4815) e, mais importante, cumpriu todos os pressupostos da regressão linear, garantindo uma maior robustez.

Diante destes resultados, o *modelo5* e o *modelo6* aparentam ser as melhores escolhas, pois equilibram um bom poder explicativo com a validação dos pressupostos estatísticos. O *modelo5*, ao transformar a variável dependente, obteve o maior R^2, indicando um ajuste superior, embora ainda apresente problemas de independência dos resíduos. Já o *modelo6*, ao considerar o consumo per capita, cumpriu todos os pressupostos dos resíduos e apresentou o menor *AIC*.

Antes de escolher o modelo final, iremos testar a capacidade de previsão dos modelos desenvolvidos.

Avaliação

Para avaliar o desempenho dos modelos, recorremos a um conjunto treino-teste na proporção 80-20. O desempenho de previsão de cada modelo é avaliado com recurso ao *MAPE*, que indica o erro percentual absoluto médio associado a cada previsão, utilizando os valores reais (*actual*) da amostra e os valores previstos pelo modelo (*predicted*) e ao *RMSE*, que indica a variância. Quanto menores forem os valores, tanto do *MAPE* como do *RMSE*, melhor será a capacidade de previsão do modelo.

Para garantir uma avaliação mais robusta e abrangente, realizámos uma previsão *out-sample*, cujo objetivo foi testar a capacidade de generalização dos modelos em dados não observados durante o treino. Essa abordagem permite verificar se os modelos são capazes de realizar previsões consistentes e precisas em cenários reais, utilizando os 20% dos dados reservados para o conjunto de teste.

Previsão *Out-Sample*

Conforme mencionado anteriormente, 80% dos dados foram utilizados para treinar os modelos, e os 20% restantes foram reservados para o conjunto de teste.

Modelos	MAPE (%)	RMSE
Modelo1	40.119	0.132
Modelo2	38.161	0.130
Modelo3	38.386	0.130
Modelo4	37.571	0.130
Modelo5	36.547	0.135
Modelo6	54.412	0.246

O modelo5 destacou-se por apresentar melhores capacidades preditivas, com um MAPE de 36.547% e um RMSE de 0.135, tal indica que este modelo tem uma boa capacidade preditiva e forte explicação de variância dos dados.

Por outro lado, o modelo6 teve o pior desempenho, com um MAPE de 54.412% e um RMSE de 0.246, os valores mais altos da tabela. Esses resultados sugerem que, apesar de o modelo6 ser um dos mais ajustados aos dados de treino, o mesmo pode ter limitações na generalização para os dados não treinados. Os valores elevados de *MAPE* e *RMSE* indicam que o modelo tem dificuldades em fazer previsões fora da amostra de treino, o que pode levar a um risco de *overfitting*.

Um fator importante a ser considerado é que o modelo6 utiliza *wfood/size* como variável alvo e embora essa abordagem tenha resolvido o problema dos pressupostos, não foi benéfica em termos preditivos. Como a variável dependente é normalizada pelo tamanho da família (*size*), a transformação pode não ser adequada para a previsão da proporção dos gastos em comida (*wfood*). Ao multiplicar a previsão pelo número de elementos do agregado familiar (*size)* para obter *wfood*, o erro de previsão pode aumentar, especialmente para famílias muito grandes. Ao estar a prever a despesa alimentar por pessoa em vez do agregado familiar no seu todo, o modelo não tem em conta que as famílias maiores tendem a gastar menos em alimentação (por pessoa) devido a fazerem compras em maiores quantidades, por exemplo.

Escolha do Modelo Final

Tendo em conta todos os resultados obtidos, o *modelo5* foi escolhido como o modelo final para previsão. Ele apresentou o maior R^2 além de possuir o menor Erro Percentual Absoluto Médio (*MAPE)*, demonstrando uma melhor capacidade preditiva e um menor erro nas previsões. Embora o *AIC* seja elevado, o seu desempenho preditivo justifica a sua escolha, garantindo um equilíbrio entre ajuste e generalização. Assim, o modelo5 foi considerado o mais adequado para representar o comportamento da variável dependente e gerar previsões mais confiáveis.

Em baixo, segue-se o gráfico de previsão do modelo5 no conjunto de teste, onde são previstas 50 observações da base de dados.

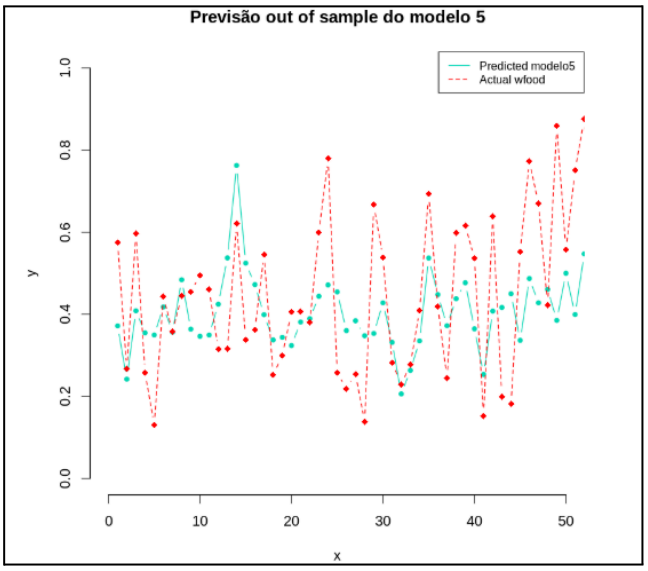

Fig. 24 - Previsão *out-sample* do Modelo 5 (50 amostras)

Interpretação

Após a escolha do melhor modelo, neste caso, o modelo5, procede-se à interpretação dos coeficientes das variáveis independentes incluídas no modelo. Os coeficientes das variáveis independentes indicam a variação da variável dependente – que, neste modelo, corresponde ao valor absoluto gasto na alimentação por agregado familiar.

Abaixo, apresenta-se o sumário do modelo5:

```
Call:
lm(formula = log(wfood * totexp) ~ log(totexp) + I(totexp^2) +
    sex + age + log(totexp):size + log(totexp):as.factor(town),
    data = budget_food_data, weights = 1/sqrt(totexp))

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.62257 -0.02437  0.00404  0.03261  0.25232

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  7.492e-01  1.330e-01   5.633 1.97e-08 ***
log(totexp)                  7.116e-01  1.558e-02  45.662  < 2e-16 ***
I(totexp^2)                 -1.293e-09  1.421e-10  -9.100  < 2e-16 ***
sex                          7.919e-02  2.592e-02   3.055  0.00227 **
age                          3.812e-03  6.831e-04   5.581 2.65e-08 ***
log(totexp):size             9.515e-03  7.359e-04  12.930  < 2e-16 ***
log(totexp):as.factor(town)1 2.102e-02  3.265e-03   6.436 1.46e-10 ***
log(totexp):as.factor(town)2 8.505e-03  2.985e-03   2.850  0.00441 **
log(totexp):as.factor(town)3 6.509e-03  3.003e-03   2.168  0.03027 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05738 on 2481 degrees of freedom
Multiple R-squared:  0.6847,    Adjusted R-squared:  0.6837
F-statistic: 673.6 on 8 and 2481 DF,  p-value: < 2.2e-16
```

Fig. 25 - Sumário do modelo5

Diante destes resultados, a próxima etapa consiste na análise individual dos coeficientes das variáveis preditoras, a fim de compreender os seus impactos sobre a variável dependente.

- **(Intercept)** -> Representa o valor esperado do gasto total em alimentação quando todas as variáveis explicativas são zero. No entanto, a sua interpretação por si só não é indicativa.
- **log(totexp)** -> Indica que, mantendo todas as outras variáveis constantes, um aumento de 1% na despesa total está associado a um aumento de aproximadamente 0.71% na despesa alimentar. Ou seja, a despesa alimentar cresce, mas a um ritmo proporcionalmente menor ao aumento das despesas totais.
- **I(totexp^2)** -> O termo quadrático sugere uma relação não linear entre a despesa total e a despesa na alimentação. O coeficiente negativo indica que, após certo ponto, o aumento da despesa total tem um efeito decrescente sobre a despesa na alimentação, confirmando a ***lei de Engel*** (a proporção da renda gasta em alimentação diminui com o aumento da renda).

- **sex** -> O coeficiente positivo sugere que indivíduos do sexo masculino gastam aproximadamente 7.9% mais em alimentação do que os do sexo feminino, mantendo todas as outras variáveis constantes.
- **age** -> Por cada ano de idade adicional associa-se um aumento de aproximadamente 0.38% no valor gasto na alimentação, mantendo-se tudo o resto constante;
- **log(totexp):size** -> O impacto das despesas totais sobre o valor absoluto gasto na alimentação aumenta conforme o número de membros familiares cresce. Em termos numéricos, para um aumento de 1% nas despesas totais e um acréscimo de 1 unidade em size, as despesas na alimentação aumentam 0.009515€.
- **log(totexp):as.factor(town)1** -> Para famílias residentes em cidades pequenas (**town1)**, o impacto no gasto alimentar após um aumento percentual nas despesas totais é **2.102% maior** do que para famílias residentes em cidades média-grandes (**town4).**
- **log(totexp):as.factor(town)2 ->** Para famílias residentes em cidades de dimensão 2 (**town2)**, o impacto no gasto alimentar após um aumento percentual nas despesas totais é **0.85% maior** do que para famílias residentes em cidades média-grandes (**town4).**
- **log(totexp):as.factor(town)3 ->** Para famílias residentes em cidades médias (**town3)**, o impacto no gasto alimentar após um aumento percentual nas despesas totais é **0.65% maior** do que para famílias residentes em cidades média-grandes (**town4).**

Ou seja, quanto menor a cidade do inquirido, maior o efeito que as despesas totais têm no gasto alimentar da sua família.

Conclusão

Neste trabalho, desenvolvemos e analisámos vários modelos de regressão linear múltipla com o objetivo de prever a percentagem da despesa total destinada à alimentação nos agregados familiares em Espanha. A partir de uma abordagem sistemática baseada na metodologia *CRISP-DM*, percorremos as etapas de compreensão do problema, exploração dos dados, transformação das variáveis e desenvolvimento de múltiplos modelos, avaliando a sua qualidade com base em métricas como R^2, *AIC* e a verificação dos pressupostos estatísticos.

Os primeiros modelos apresentaram limitações devido a problemas como heterocedasticidade e autocorrelação dos resíduos, comprometendo a robustez das previsões. Para mitigar essas questões, aplicámos transformações e ponderações nos dados, o que levou à melhoria do desempenho preditivo dos modelos subsequentes. Em particular, o **modelo 5**, que considerou a transformação logarítmica da variável dependente, destacou-se pelo melhor equilíbrio entre capacidade explicativa (R^2) e desempenho preditivo em dados fora da amostra, tornando-se a escolha final.

A análise dos coeficientes do modelo final permitiu identificar os principais fatores que influenciam a proporção do orçamento familiar destinada à alimentação. Em particular, confirmámos a validade da *Lei de Engel*, que diz que agregados familiares com maiores rendimentos tendem a destinar uma menor proporção do orçamento à alimentação, além de outras relações estatisticamente significativas entre variáveis demográficas e económicas.

Por fim, este estudo demonstrou a importância da escolha criteriosa dos modelos e da verificação dos pressupostos estatísticos na construção de modelos preditivos fiáveis. Apesar dos avanços obtidos, futuros estudos poderiam explorar técnicas mais avançadas, como modelos não lineares ou métodos de machine learning, para aprimorar ainda mais a precisão das previsões.