UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Identificação de padrões alimentares na dieta de doentes com Degenerescência Macular Relacionada com a Idade

Ana Raquel Duque Gonçalves Fernandes

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Marília Cristina de Sousa Antunes
Doutora Sandrina Gonçalves Nunes

2017

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Identificação de padrões alimentares na dieta de doentes com Degenerescência Macular Relacionada com a Idade

Ana Raquel Duque Gonçalves Fernandes

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Marília Cristina de Sousa Antunes
Doutora Sandrina Gonçalves Nunes

2017

Agradecimentos

O espaço limitado desta secção de agradecimentos, seguramente, não me permite agradecer devidamente a todas as pessoas que, ao longo do Mestrado em Bioestatística, me ajudaram, direta ou indiretamente, a cumprir os meus objetivos e a realizar mais esta etapa da minha formação académica. Deixo então apenas algumas palavras que transmitem o meu profundo e eterno sentimento de reconhecido agradecimento.

À Professora Doutora Marília Antunes, pela sua orientação, total apoio, disponibilidade, pelo conhecimento que transmitiu, pelas opiniões e críticas e total colaboração no solucionar de dúvidas e problemas que foram surgindo ao longo da realização deste trabalho. Quero inclusivamente agradecer a força e as oportunidades que me proporcionou, que espero um dia poder vir a retribuir! O seu apoio foi determinante na elaboração deste Projeto. Obrigada professora!

À Doutora Sandrina Nunes, por me ter recebido na Associação para Investigação Biomédica em Luz e Imagem (AIBILI) e por me ter proporcionado as condições necessárias para a elaboração do meu Projeto, permitindo a minha integração nesta Instituição. Agradeço também a sua disponibilidade, colaboração e cuidado que dedicou a este Projeto.

Não deixando de agradecer à AIBILI, pela oportunidade e a toda a equipa. Um obrigada muito especial à Dalila, Miguel e Liliana, que me acompanharam nesta fase tão turbulenta, me ensinaram, me motivaram e estiveram sempre disponíveis para me ajudar.

Expresso também a minha gratidão a todos os participantes do estudo que, embora no anonimato, prestaram uma contribuição fundamental para que este estudo fosse possível.

A todos os professores do Mestrado, agradeço a oportunidade e o privilégio que tive em frequentar este Mestrado que em muito contribuiu para o enriquecimento da minha formação académica e científica. Um agradecimento especial à Professora Lisete, pelo profissionalismo, apoio, disponibilidade e simpatia.

Aos meus colegas de Mestrado pelo companheirismo e boa disposição, que me ajudaram a transitar para este novo Mundo da Estatística. Uma referência especial ao Luís, sempre prestável e amigo, não há palavras para descrever o *template* do LaTeX.

Ao Miguel, que as aulas de Programação nos apresentaram oficialmente, obrigada por tudo, foste

das pessoas mais importantes desta fase turbulenta!

Ao Ricardo Martins, por me teres apoiado, com tanta paciência, e por teres estado presente numa grande parte da minha vida e impulsionado para ser melhor a cada dia.

À Lídia Grilo, pelo apoio não só neste Mestrado, como também no anterior. Obrigada por me teres ouvido, compreendido e aconselhado, principalmente nestes últimos meses.

Aos meus amigos Fábio Duarte e Fábio Gomes, entre outros que não menciono o nome mas sabem quem são, pela compreensão, carinho, confiança e apoio. Foram imprescindíveis nesta fase repleta de momentos difíceis.

À Inês Martins, a minha madrinha académica, que levo sempre comigo. Do início ao fim deste percurso académico me acompanhaste, e sempre serás uma inspiração para mim.

À Patrícia Serra, não tenho palavras para descrever tudo o que tens feito por mim. Obrigada por teres entrado na minha vida e por seres quem és. Obrigada também por me teres trazido mais um grupo de amigas fantásticas. (Paciência!)

Por último, tendo consciência que sozinha nada disto seria possível, dirijo um especial agradecimento aos meus pais, por serem modelos de coragem, pelo seu apoio incondicional, sacrifício, incentivo, amizade e paciência (alguma) demonstrados e total ajuda na superação dos obstáculos que ao longo da minha derradeira caminhada foram surgindo. Ao meu irmão, por ser o meu orgulho eterno e aos meus avós, que sempre me acarinharam e apoiaram sem reservas. A eles dedico este trabalho.

Resumo

A degenerescência macular da idade (DMI) é a principal causa de cegueira irreversível entre as pessoas com idade igual ou superior a 55 anos nos países desenvolvidos (Shaw et al., 2016; Lim et al., 2012). Atualmente, os tratamentos de DMI têm um custo elevado e restringem-se apenas ao retardamento da progressão da forma neovascular da doença. Vários estudos sugeriram que alguns aspetos da dieta podem influenciar o risco de desenvolvimento e/ou progressão da doença, no entanto, os resultados obtidos foram maioritariamente baseados em análises do consumo individual de nutrientes e alimentos ou grupos de alimentos (Age-Related Eye Disease Study Research Group et al., 2008; Tan et al., 2009; Chong et al., 2009). A avaliação de padrões dietéticos tem surgido, na área da epidemiologia nutricional, como uma ferramenta mais informativa e preditiva do risco da doença, uma vez que permite estudar a dieta como um todo (Hu, 2002). Apenas dois artigos reportaram a avaliação de associações entre padrões alimentares e o risco de DMI (Chiu et al., 2014; Islam et al., 2014), contudo, ainda nenhum estudo foi realizado para a população portuguesa. Este projeto pretendeu, deste modo, identificar os principais padrões alimentares de uma subpopulação com idade superior a 55 anos da Lousã, uma região rural do interior de Portugal, e analisar a respetiva associação com o risco de DMI.

Para tal, foram analisados os dados demográficos, de história médica geral e oftalmológica referentes a uma amostra de 1000 participantes da Unidade de Saúde Familiar da Lousã, com uma proporção de 1:1 de sujeitos com e sem DMI, disponíveis na base de dados construída no estudo transversal *Coimbra Eye Study* e também os dados obtidos pela realização de um questionário semiquantitativo de frequência alimentar, previamente validado para a população portuguesa.

Inicialmente, para a determinação dos principais padrões alimentares da população em estudo, foi utilizada a metodologia de análise em componentes principais (ACP), e confirmados os respetivos resultados com análise classificatória (AC). Seguidamente, procedeu-se à construção de modelos de regressão logística binária, onde a variável resposta consistiu na presença/ausência de DMI. As variáveis explicativas consideradas foram os padrões alimentares retidos com ACP e outros fatores de risco potenciais da doença, como a idade, sexo, escolaridade, índice de massa corporal (IMC), perímetro abdominal, hábitos tabágicos, atividade física, consumo energético total e alcoólico, diabetes, hipertensão e dislipidémia. Uma vez que este projeto teve como foco analisar especificamente os efeitos dos padrões alimentares com a doença, procurou-se diminuir o confundimento residual através do desenvolvimento e aplicação de um método que permitiu simultaneamente a seleção das variáveis de confundimento e da forma funcional das covariáveis contínuas.

Neste estudo foram identificados três padrões alimentares: um padrão que refletiu uma dieta tradicional, típica da região interior portuguesa, um padrão saudável e um padrão menos saudável, que se distinguiu dos anteriores pelo elevado consumo de alimentos de fácil preparação ou petiscos. Foram encontradas associações entre o padrão tradicional e o padrão saudável e o risco de DMI, porém apenas ao nível de significância de $\alpha = 0.1$. O padrão tradicional revelou estar associado positivamente com o risco de DMI, enquanto que o padrão saudável revelou ter um papel protetor no desenvolvimento da doença. O padrão menos saudável não demonstrou qualquer associação significativa com o risco de DMI.

Os resultados aqui obtidos suportam, assim, o papel importante da dieta no desenvolvimento de DMI, à semelhança do ocorrido nos outros dois estudos realizados para outras populações. Tal poderá providenciar uma melhor compreensão da relação entre a doença e as práticas alimentares na população estudada e poderá fornecer informação adicional para um maior apoio à educação, aconselhamento e intervenção nutricional. Neste estudo, concluiu-se que uma dieta rica em iogurte, peixe, fruta, hortícolas, salada e sopa é mais aconselhável para prevenção da doença quando comparada a uma dieta abundante em carnes vermelhas, bacalhau, acompanhamentos (arroz, massa, batatas cozidas ou assadas), pão branco ou integral, tostas e broa, azeite, bebidas alcoólicas e café.

Palavras-Chave: Degenerescência macular da idade, Padrões alimentares, Análise em componentes principais, Modelos de regressão logística binária, Confundimento residual

Abstract

Age-related macular degeneration (AMD) is the major cause of irreversible blindness among people aged 55 or over in developed countries(Shaw et al., 2016; Lim et al., 2012). Currently, clinical treatments for AMD are costly and are limited to arresting the neovascular type of the disease. Several studies have suggested that some aspects of diet could influence the risk of disease development and/or progression, although the results have largely been based on analysis evaluating individual nutrient, food or food groups (Age-Related Eye Disease Study Research Group et al., 2008; Tan et al., 2009; Chong et al., 2009). Dietary pattern analysis has emerged in the field of nutritional epidemiology as a more informative and predictive tool of disease risk, since it allows to study a diet as a whole (Hu, 2002) to understand how it can be optimized to promote health. Only two articles have examined the relationship between dietary patterns and the risk of AMD (Chiu et al., 2014; Islam et al., 2014), however no studies have yet been reported for the Portuguese population. Thus, initially, this project aimed to identify major food patterns of a subpopulation over 55 years of Lousã, a rural interior area in Portugal, and, subsequently, to evaluate potential associations with the risk of AMD.

According to the objectives set, one analysed demographic, general medical and ophthalmological data collected from a sample of 1000 participants from the Family Health Unit of Lousã, with a ratio of 1:1 of subjects with and without AMD, available in the database constructed in the cross-sectional study Coimbra Eye Study and also the data obtained by administering a semi-quantitative food frequency questionnaire previously validated for the Portuguese population.

Principal components analysis (PCA) was used to determine the major dietary patterns of the study population, and the results obtained were confirmed with clustering analysis (CA). Afterwards, binary logistic regression models were constructed, where the response variable was the presence/absence of AMD. The explanatory variables considered were food patterns retained with PCA and other potential risk factors of the disease, such as age, sex, formal education, body mass index (BMI), abdominal circumference, smoking habits, physical activity, total energy intake and alcohol intake, diabetes, hypertension and dyslipidemia. Since the main focus of this project was to comprehend the effects of dietary patterns on the disease, it was sought to reduce residual confounding through the development and application of a method that simultaneously allowed the selection of confounding variables and the functional forms of continuous covariates.

Three dietary patterns were identified: a pattern that reflected a traditional diet typical of the Portuguese interior region, a healthy pattern and a less healthy pattern, which was distinguished from

previous ones by the high consumption of easily prepared foods or snacks. Associations were found between the traditional pattern and the healthy pattern and the risk of AMD, but only at the significance level of $\alpha = 0.1$. The traditional pattern seems to potentiate the risk of AMD, while the healthy pattern seems to play a protective role in the development of AMD. The less healthy pattern did not shown any significant association with the risk of AMD.

The results obtained here support the important role of diet in the development of AMD, as in the other two studies performed for other populations. This may provide a better understanding of the relationship between disease and dietary practices in the population studied and may provide additional information for further support to education, counseling and nutritional intervention. In this study, it was concluded that a diet rich in yogurt, fish, fruit, vegetables, salads and soup is more advisable for disease prevention when compared to a diet rich in red meat, cod, side dishes (rice, pasta, cooked or roasted potatoes), white or whole grain bread, toasted bread, olive oil, alcoholic beverages and coffee.

Keywords: Age-related macular degeneration, Food patterns, Principal component analysis, Binary logistic regression models, Residual confounding

Conteúdo

Lista de siglas e abreviaturas

ACP Análise em Componentes Principais

AC Análise Classificatória

AFE Análise Fatorial Exploratória

AIBILI Associação para Investigação Biomédica em Luz e Imagem

AIC *Akaike information criterion*

aMeDi *Alternate Mediterranean Diet*

ANOVA *Analysis of Variance*

AREDS *Age-Related Eye Disease Study*

CAREDS *Carotenoids Age-Related Eye Disease Study*

CES Comissão de Ética para a Saúde

CFH Fator H do Complemento

CONSORT *Consolidated Standards of Reporting Trials*

CORC *Coimbra Ophthalmology Reading Center*

CP Componente Principal

CRAN *Comprehensive R Archive Network*

DGS Direção-Geral da Saúde

DHA Ácido Docosa-Hexaenoico

DMI Degenerescência Macular da Idade

EMV Estimador de Máxima Verosimilhança

EPA Ácido Eicosapentaenoico

GVIF *Generalized Variance Inflation Factor*

ICGS *International Classification and Grading System*

ID Número de identificação do participante

IMC Índice de Massa Corporal

INSA Instituto Nacional de Saúde Doutor Ricardo Jorge

LOWESS *Locally Weighted Scatterplot Smoothing*

mHEI *Healthy Eating Index* modificado

MLG Modelos Lineares Generalizados

POLA *Pathologies Oculaires Liées à l'Age*

RCS *Restricted Cubic Splines*

RRR *Reduced Rank Regression*

RV Razão de Verosimilhanças

TLC Teorema Limite Central

VIF *Variance Inflation Factor*

OR *Odds-Ratio*

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Degenerescência macular da idade

A degenerescência macular da idade (DMI) é uma doença de evolução crónica que afeta a retina central. Esta consiste na maior causa de cegueira irreversível entre a população com idade igual ou superior a 55 anos em países desenvolvidos, afetando dezenas de milhões de indivíduos dessa faixa etária por todo o Mundo (Tomany et al., 2004; Shaw et al., 2016; Lim et al., 2012; Friedman et al., 2004). Em Portugal, estima-se que cerca de 12.5% da população seja afetada por DMI (Direção-Geral da Saúde, 2016).

A DMI caracteriza-se essencialmente pela presença de pequenos depósitos amarelos de lípidos e proteínas (*soft drusen*) na área da região central da retina, denominada de mácula (Jager et al., 2008). Segundo a classificação internacional da DMI definida em *International Age-Related Macular Epidemiological Study Group Classification* (Bird et al., 1995), consideram-se duas fases no desenvolvimento da doença. A fase precoce (DMI precoce), geralmente assintomática, que contribui em cerca de 80% dos casos de DMI, que pode evoluir para a fase denominada tardia (DMI tardia ou avançada).

A fase tardia compreende duas formas: a forma atrófica (DMI não-exsudativa ou atrofia geográfica) (Sunness, 1999); e a forma exsudativa (DMI exsudativa ou neovascularização coroidal), a mais agressiva, contando com aproximadamente 90% dos casos de perda visual severa ou perda irreversível da visão resultante de DMI (Correia, 2002; Shaw et al., 2016).

A prevalência de DMI, dada a sua natureza crónica, tem aumento acentuado com a idade. Face ao atual envelhecimento da população e aumento da esperança média de vida, tal traduz-se num grave problema de saúde pública, com implicações não só na qualidade de vida dos indivíduos desta faixa etária que apresentam as formas avançadas desta doença, mas também a nível económico, com o crescente consumo de recursos de saúde (Chiu et al., 2007b; Klein, 2007).

Apesar do grande impacto desta doença na população e do esforço por parte da comunidade científica, dada a sua patogénese ser resultante de complexas interações entre fatores de risco genéticos e ambientais (Rosenfeld et al., 2006), a sua etiologia ainda não se encontra completamente esclarecida, o que levanta barreiras na criação de terapêuticas eficazes para a respetiva prevenção e tratamento (Chiu et al., 2007b). De facto, até ao momento existem apenas formas de tratamento efetivas para evitar a progressão da forma exsudativa de DMI, consistindo em injeções periódicas intraoculares de

medicamentos anti-angiogénicos (exemplo: Ranibizumab, Aflibercept), com altos custos associados (Rosenfeld et al., 2006).

Os factos apresentados evidenciam a importância da adoção de medidas preventivas através da intervenção sobre os fatores de risco modificáveis conhecidos de modo a reduzir o impacto desta doença na população.

Determinadas comorbilidades e comportamentos adjacentes ao estilo de vida praticado pelo indivíduo, como o sedentarismo, hábitos tabágicos, elevado colesterol, obesidade, exposição solar, dieta, entre outros (Age-Related Eye Disease Study Research Group et al., 2008; Curcio et al., 2009; Hawkins et al., 1999; Khotcharrat et al., 2015; Klein et al., 2004; Kiernan et al., 2010; Meyers et al., 2015; Seddon et al., 2006), constituem fatores de risco já bem estabelecidos na literatura por exercerem influência não só no risco de desenvolvimento e/ou progressão da DMI, como também de outras patologias crónicas, como a aterosclerose, cancro, Alzheimer, Parkinson, entre outras.

Muitos destes fatores de risco identificados têm efeitos bem documentados num mecanismo biológico natural denominado stress oxidativo e consequente inflamação, mecanismo este que tem um papel conhecido na patogénese das doenças relacionadas com a idade (Seddon et al., 2006; Schick et al., 2016; Millen et al., 2015; Binder et al., 2002; Brewer, 2007; Beatty et al., 2000; Hollyfield et al., 2008; Kim et al., 2015). A manipulação do equilíbrio de antioxidantes no organismo através da dieta ou suplementação surgiu, assim, como uma potencial estratégia com vista a prevenir o desenvolvimento e a progressão de DMI de maneira prática e custo-efetiva (Chiu and Taylor, 2007; Sin et al., 2013).

O *Age-Related Eye Disease Study* (AREDS) do *National Eye Institute of the National Institutes of Health* (Age-Related Eye Disease Study Research Group et al., 2008), um ensaio clínico de larga escala aleatorizado, interventivo e multicêntrico, com duração média de 6.3 anos, foi o primeiro a demonstrar o benefício substancial da suplementação com antioxidantes e minerais (altas doses de vitamina C e E, β-caroteno, e zinco) na redução na progressão para a forma avançada de DMI, em pessoas com idade superior a 55 anos (risco de DMI reduzido em 28%) (Sin et al., 2013).

Outros estudos dietéticos têm sido publicados, na sua maioria com vista a analisar a relação entre DMI e o consumo de micronutrientes e macronutrientes, assim como a sua associação com o consumo de alimentos e grupos alimentares a nível individual. O benefício da suplementação com xantofilas maculares (luteína e zeaxantina) foi demonstrado no estudo AREDS2, mas não da suplementação com ómega-3, como verificado noutro estudo (Merle et al., 2011). As dietas com alto consumo de peixe têm sido também associadas com uma diminuição do risco de DMI (Tan et al., 2009; Swenor et al., 2010), contrariamente às dietas ricas em gorduras *trans*, carnes vermelhas, álcool e alimentos com alto índice glicémico, que revelaram ter um efeito prejudicial (Chong et al., 2009, 2008; Adams et al., 2012; Chiu et al., 2007a). Contudo, embora os resultados obtidos sugiram que alguns aspetos da dieta podem alterar o risco da doença, estes encontram-se frequentemente inconsistentes ao longo de vários estudos. Adicionalmente, e não menos importante, estes estudos não consideram as possíveis interações entre os diversos constituintes dos alimentos e a consequente influência no risco da doença (Chiu et al., 2009).

1.2 Padrões alimentares e relação com a doença

Na área da epidemiologia nutricional, a análise dos padrões alimentares tem surgido como uma abordagem alternativa para estudar a relação entre a dieta e o risco de doenças crónicas. Ao invés de estudar os alimentos ou nutrientes individualmente, esta abordagem permite analisar os efeitos da dieta como um todo, tendo em conta a composição relativa dos grupos alimentares e suas interações e correlação, podendo ser, deste modo, mais preditiva e informativa do risco de doença (Hu, 2002; Chiu et al., 2014; Islam et al., 2014).

Vários métodos se encontram descritos com a finalidade de derivar os padrões alimentares numa população, sendo divididos em métodos *a priori* e métodos *a posteriori*.

Os métodos *a priori* foram os primeiros a ser utilizados com esse objetivo, recorrendo a índices dietéticos ou *scores* baseados em critérios ou padrões alimentares predefinidos pelo investigador, de modo a avaliar o grau de adesão dum participante a esse regime (Kant, 1996, 2004).

Os métodos *a posteriori* derivam os padrões alimentares empiricamente, tendo como base o consumo dietético observado, refletindo, consequentemente, as dietas selecionadas e consumidas pelos indivíduos da população em estudo (Wirfält et al., 2013). Estes métodos, originalmente desenvolvidos como ferramentas estatísticas para a redução da dimensionalidade de grandes conjuntos de dados com elevado número de variáveis, têm sido aplicados nos últimos 30 anos na área da epidemiologia nutricional. Entre eles, encontram-se reportados a análise em componentes principais (ACP), a análise fatorial exploratória (AFE), a análise classificatória (AC) e, mais recentemente, a regressão por posto reduzido (*Reduced Rank Regression* ou RRR).

Neste contexto, surgiu a hipótese de modelar o risco de desenvolvimento de DMI pela dieta no seu geral, sem recurso a suplementos, que apesar de se revelar atrativa, ainda se encontra relativamente pouco descrita na literatura disponível (Mares et al., 2011; Gopinath et al., 2015; Chiu et al., 2014; Islam et al., 2014).

O estudo *Dietary Patterns and Their Associations with Age-related Macular Degeneration* (Islam et al., 2014) foi dos primeiros a investigar a associação do risco de DMI com padrões alimentares identificados na população selecionada do estudo *Melbourne Collaborative Cohort Study*, um estudo prospetivo conduzido em Austrália entre 1990 e 1994. Os resultados obtidos demonstraram evidência significativa de associação entre um padrão rico em frutas, vegetais, frango e frutos secos e pobre em carnes vermelhas com uma menor prevalência de formas avançadas de DMI.

Neste estudo foi utilizado o método de análise de componentes principais para a identificação de padrões alimentares, seguido pela construção de modelos de regressão logística binária para investigar a associação do risco de DMI com os padrões alimentares obtidos anteriormente. Com objetivos e metodologias semelhantes, encontram-se descritos na literatura outros estudos, entre eles o estudo *The relationship of major American dietary patterns to age-related macular degeneration* (Chiu et al., 2014), tendo como base a construção de padrões alimentares de diferentes populações e a análise do respetivo impacto no risco de DMI.

Relativamente à população portuguesa, encontra-se reportado o estudo *Nutritional and lifestyle*

risk factors in AMD (NCT01715870[a]). *The Coimbra Eye Study* (NCT01298674[a]), com o objetivo de comparar os padrões alimentares e estilos de vida entre indivíduos com e sem DMI. Foi utilizada uma escala ordinal de adesão à dieta mediterrânica, o mediSCORE, obtida pela soma de 9 variáveis indicadoras de consumo pelos principais grupos alimentares (vegetais, legumes, frutas, cereais, peixe, carne, produtos láteos, álcool e rácio de monolípidos para gordura saturada). No entanto, este estudo envolve um método *a priori*, sendo limitado pelo conhecimento prévio da relação entre a dieta e a doença, e construído segundo recomendações dietéticas, não identificando, portanto, os padrões alimentares empíricos desta população.

Neste projeto pretende-se identificar os principais padrões alimentares da população da Lousã, para subsequente avaliação do impacto desses padrões no risco de DMI. Com esse objetivo, foram aplicadas duas técnicas de análise multivariada distintas para extrair os principais padrões alimentares, a análise em componentes principais e a análise classificatória, e posteriormente construído um modelo de regressão logística ajustado às covariáveis consideradas clinicamente relevantes para a doença, à semelhança dos estudos acima referidos.

No capítulo 2 encontram-se descritos alguns procedimentos referentes ao delineamento experimental deste projeto, como a seleção da amostra, inclusão de participantes, recolha de dados e exclusão de participantes. O capítulo 3 foca-se nos procedimentos respeitantes à construção e processamento da base de dados, fornecendo, inclusivamente, informação sobre a transformação, tratamento de valores omissos e categorização realizados sobre as variáveis da base de dados original. O capítulo 4 aborda os principais métodos de análise estatística utilizados, entre eles a análise em componentes principais, a análise classificatória e a análise de regressão logística múltipla, introduzindo as ideias teóricas básicas e alguns procedimentos a realizar. No capítulo 5 são apresentados os resultados derivados da aplicação dessas diferentes metodologias estatísticas. De acordo com esses resultados, a informação inerente aos dados é discutida no capítulo 6, onde são também comparados os resultados obtidos noutros estudos, e apresentadas limitações adjacentes à metodologia utilizada neste projeto. Finalmente, é providenciada uma conclusão do trabalho aqui realizado no capítulo 7.

[a]Identificador do estudo na base de dados pública ClinicalTrial.Gov, **http://www.clinicaltrials.gov**.

2 Delineamento experimental

2.1 Seleção da amostra

A amostra utilizada no âmbito deste projeto foi selecionada do estudo *Coimbra Eye Study*. Este é um estudo observacional e transversal, com o objetivo de analisar a prevalência das formas precoces e avançadas de DMI, que incluiu indivíduos com idade superior a 55 anos de unidades de saúde de dois locais no centro de Portugal - um na zona costeira (Mira) e outro a 70 km do mar (Lousã).

Para o presente projeto, foram selecionados desse estudo 1000 participantes da Unidade de Saúde Familiar da Lousã. A seleção dos indivíduos consistiu na inclusão de 500 participantes que apresentavam DMI (todos os sujeitos com fases avançadas da doença foram incluídos) e de 500 participantes sem DMI, sempre que possível emparelhados por sexo e idade.

A recolha de dados envolveu uma única visita, onde os sujeitos foram convidados a responder a um questionário validado de hábitos alimentares e de estilos de vida (Apêndice A), numa entrevista conduzida por um nutricionista devidamente treinado. Caso confirmada a disponibilidade dos indivíduos para participação no estudo, estes assinaram o Consentimento Informado, onde foi explicado o objetivo do estudo e autorizado o uso dos dados, sempre assegurando a confidencialidade da identificação do participante.

O processo de recrutamento teve a duração de 12 meses e os indivíduos foram contactados para uma visita nesse período de recrutamento. Caso não pudessem comparecer nessa data, a visita seria remarcada.

O estudo foi submetido e aprovado pela Comissão de Ética para a Saúde (CES) da Associação para Investigação Biomédica em Luz e Imagem (AIBILI), estando registado na base de dados pública **ClinicalTrials.Gov** com o número NCT01715870.

2.2 Inclusão de participantes

A partir da base de dados do estudo *Coimbra Eye Study*, foi construída uma base de dados respeitante apenas à população da Lousã. Todos os doentes desta base de dados realizaram uma avaliação oftalmológica bilateral completa, com avaliação da melhor acuidade visual corrigida, biomicroscopia do segmento anterior, tonometria e fotografia a cores do fundo midriática digital (Topcon®TRC-50EX,

Topcon Corp, Tóquio, Japão). As fotografias do fundo do olho foram analisadas num centro de classificação de imagens oftalmológicas (*Coimbra Ophthalmology Reading Center*, CORC - AIBILI) e uma análise diferencial para as lesões de DMI foi conduzida por dois oftalmologistas séniores, independentes e certificados, usando a *International Classification and Grading System (ICGS) for ARM and AMD* (Bird et al., 1995) e a classificação de Roterdão (Vingerling et al., 1995a). O protocolo completo do estudo *Coimbra Eye Study* encontra-se detalhado com mais pormenor na literatura (Cachulo et al., 2015).

2.3 Recolha de dados

Os dados demográficos, de história médica geral e oftalmológica utilizados neste projeto foram recolhidos do estudo *Coimbra Eye Study*, conjuntamente com os dados recolhidos do questionário referentes à avaliação dos hábitos alimentares e consumo de nutrientes.

O questionário semiquantitativo de frequência alimentar administrado aos participantes teve como finalidade a recolha de informações quantitativas do consumo alimentar referentes ao período de 12 meses anterior à sua realização. Este questionário foi desenvolvido e validado para a população Portuguesa no ano 2000 pelo Serviço de Higiene e Epidemiologia da Faculdade de Medicina da Universidade do Porto (Lopes et al., 2006), tendo sido também medida a sua reprodutibilidade (Lopes, 2000), e encontra-se disponível em **http://higiene.med.uc.pt/freq.php**.

O questionário, disponível no Apêndice A, foi preenchido por entrevistadores especialmente treinados, tendo o preenchimento sido feito na entrevista previamente marcada com os indivíduos incluídos no estudo.

Na primeira parte do questionário foram recolhidos os dados demográficos, antropométricos (altura, peso e perímetro abdominal), educacionais, associados ao estilo de vida (exercício físico e hábitos tabágicos), anos de escolaridade, comorbilidades e hábitos alimentares. O entrevistador procedeu à medida do peso, altura (para o cálculo do índice de massa corporal, IMC) e perímetro abdominal.

Na segunda parte do questionário foram recolhidos dados semiquantitativos da frequência de ingestão de alimentos. A estrutura desta parte do inquérito é a seguinte: uma lista de 86 alimentos ou grupos de alimentos; uma secção fechada com nove categorias de frequências de consumo a variar entre "Nunca ou menos de uma vez por mês"a "Seis ou mais vezes por dia"; uma secção com porções médias padrão predeterminadas; e finalmente, por uma secção aberta para o registo de outros alimentos não referenciados e consumidos com uma frequência de pelo menos uma vez por semana. Para cada item alimentar, os participantes reportaram, assim, a frequência do seu consumo no último ano, o tamanho da porção e se este consumo foi ou não sazonal.

Os 86 itens alimentares do questionário foram obtidos pelo agrupamento de alimentos de acordo com a similaridade da composição nutricional (Lopes et al., 2006). Estes grupos têm em conta a contribuição dos alimentos para a variação inter-pessoal no consumo alimentar, particularmente no que respeita à energia total, proteína, gordura total, hidratos de carbono, colesterol, fibra alimentar, vitaminas A, C, D e E, carotenoides, cálcio e etanol.

A composição quantitativa em macronutrientes e micronutrientes dos alimentos consumidos pelos participantes foi obtida usando a Tabela da Composição dos Alimentos Portugueses do Instituto Nacional

de Saúde Doutor Ricardo Jorge (INSA), publicada em 2007 (INSA, 2006). A Tabela da Composição dos Alimentos encontra-se estruturada por grupo alimentar e, dentro de cada um desses grupos, os produtos estão distribuídos por ordem alfabética e ordenados dos alimentos em natureza/crus para os cozinhados e processados. Na macro de **Excel** disponibilizada pela AIBILI, utilizada para a determinação da composição nutricional, a cada item alimentar foi associado um determinado produto da Tabela da Composição dos Alimentos, sob uma determinada forma de confeção (cru, processado ou cozinhado com outros ingredientes também, como azeite, alho, entre outros). Na tabela 7.1 do Apêndice B encontra-se a descrição dos itens alimentares do inquérito, a descrição segundo a Tabela da Composição de Alimentos do produto correspondente e o código alfanumérico associado (que permite a rastreabilidade da informação). Na tabela 7.2 do Apêndice B, encontram-se as quantidades dos nutrientes selecionados para análise neste projeto (energia total, total hidratos de carbono disponíveis, monossacarídeos+dissacarídeos, oligossacarídeos, amido, fibra alimentar, gordura total, ácidos gordos monoinsaturados, polinsaturados, saturados, *trans* e ácido linoleico, colesterol, retinol, vitamina C, alfa-tocoferol, carotenos, cálcio, ferro, magnésio, zinco e proteínas), por 100g da parte edível do produto associado ao item alimentar (excepto para as bebidas alcoólicas; valores por 100ml da parte edível).

2.4 Exclusão de participantes

Dos 1000 indivíduos selecionados inicialmente para o estudo foram excluídos 15 participantes (1.5%): um participante (0.01%) que realizou o questionário duas vezes, sendo introduzido na base de dados com um número de identificação (ID) diferente e posteriormente excluído um dos questionários (ID=1554); e 14 participantes (1.4%) para os quais o diagnóstico de DMI não foi possível por falta de qualidade fotográfica ou lesões obscuras (ID=8, 464, 1248, 1274, 1345, 1346, 1393, 1404, 1445, 1446, 1479, 1493, 1540 e 1546).

Ao contrário de outros estudos, que excluíram os doentes com diabetes pelo facto de este diagnóstico ter o potencial para uma mudança dietética (Islam et al., 2014), neste estudo optou-se por não excluir esses participantes por decisão do investigador principal. Tal deveu-se ao facto de o número de diabéticos no grupo com DMI ($n = 106$) ser semelhante ao do grupo sem DMI ($n = 155$), o que permite minimizar um potencial viés que pudesse existir nos padrões alimentares. Por outro lado, foi verificado que os 261 participantes diabéticos só apresentavam diferença significativa no consumo de açúcar comparativamente aos participantes não diabéticos, sendo que para todos os restantes alimentos não existiu evidência de diferença estatisticamente significativa. Consequentemente, a inclusão dos indivíduos diabéticos na amostra não produz nenhum viés ao nível dos padrões de consumo alimentar obtidos. Por fim, é também importante salientar que a diabetes só é diagnosticada quando já existem sinais ou suspeitas, o que poderá resultar na subestimação do número real de diabéticos no estudo, pelo que quaisquer resultados obtidos para os diabéticos nunca poderiam ser generalizados à população diabética.

Dadas as limitações deste tipo de questionário na medição de quantidades absolutas, os participantes com consumos energéticos inferiores ao percentil 1 e superiores ao percentil 99 das distribuições para cada sexo foram eliminados (n=18) (ver figura 2.1), dado apresentarem pouca probabilidade de terem reportado um consumo usual real, o que se poderia traduzir numa distorção dos padrões alimentares caso a sua inclusão se verificasse (ID=28, 29, 38, 42, 120, 473, 714, 723, 955, 1217, 1478, 1487, 1538, 1987, 2011, 2012, 2067 e 2245) (Islam et al., 2014). Os percentis 1 e 99 para as mulheres foram 775.99 e 3647.17 e para os homens 960.52 e 3924.92.

No Apêndice C encontra-se disponível o diagrama de fluxo CONSORT (*Consolidated Standards of Reporting Trials*) dos participantes do estudo selecionados e incluídos na análise (figura 7.1).

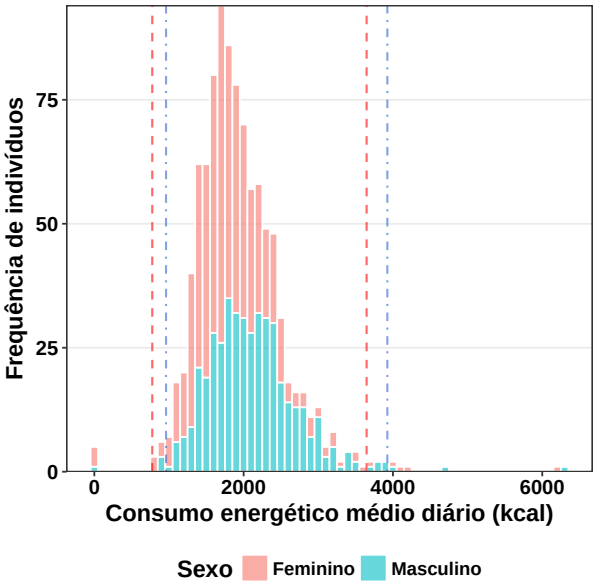


Figura 2.1: Histograma da distribuição do consumo energético médio diário (kcal) por sexo. As linhas tracejadas representam o percentil 1 e 9 para o sexo feminino e masculino (cores rosa e azul, respetivamente).

3 Base de dados

3.1 Construção da base de dados

Os inquéritos alimentares (em papel) preenchidos pelos nutricionistas foram primeiramente verificados, de modo a confirmar se se encontravam completos. Seguidamente foram inseridos manualmente numa macro de **Excel** previamente validada pela AIBILI. Esta macro realiza automaticamente a conversão dos alimentos nos respetivos macronutrientes e micronutrientes, segundo a base de dados nutricional por alimento do INSA (INSA, 2006). Esses dados foram então exportados para uma base de dados em **Excel** e importados para **Stata**.

Antes do início da análise dos dados, a base de dados foi submetida a um pré-tratamento, onde se alteraram e uniformizaram os nomes das variáveis e se codificaram uniformemente as variáveis categóricas; e foram geradas novas variáveis a partir da transformação de variáveis presentes na base de dados original, que se encontram descritas nas próximas secções. Seguiu-se uma verificação exaustiva de inconsistências, particularmente nas variáveis candidatas à análise, que por vezes implicou a consulta dos questionários em formato papel e das macros de **Excel**, de forma a detetar e corrigir valores inseridos incorretamente (erros de preenchimento ou de exportação para o **Excel**). A organização e validação da base de dados foi desenvolvida na AIBILI, responsável pela condução do estudo e centralização dos dados.

Neste capítulo pretende-se descrever algumas das variáveis da base de dados e dar informação sobre a obtenção de novas variáveis consideradas importantes para o presente estudo.

3.2 Variáveis referentes ao consumo alimentar diário

Os valores das variáveis referentes à frequência alimentar diária e porções médias de consumo foram transformados em valores numéricos, como descrito na tabela 7.3 do Apêndice B.

Foram então geradas novas variáveis de frequência alimentar (ver tabela 7.4, no Apêndice B) calculadas para os valores de consumo médio diário ajustados para o tamanho da porção, de modo a obter um valor em gramas (g) ou mililitros (ml) por tipo de alimento. Foi também incluído um fator de variação sazonal para alimentos consumidos em épocas específicas (0.25 foi considerada a sazonalidade média de três meses). A frequência alimentar ajustada foi assim calculada através da multiplicação do fator de frequência alimentar diária pelo fator de porção e pela quantidade (valor da porção média

padrão em g ou ml para cada alimento, como descrito na tabela 7.4 no Apêndice B). Caso indicado pelo participante, a frequência alimentar diária ajustada foi finalmente ponderada pelo fator de variação sazonal através da respetiva multiplicação por 0.25.

Com o objetivo de minimizar a variabilidade intra-pessoal no consumo de alimentos individuais, vários artigos respeitantes à construção de padrões alimentares reportam a classificação do elevado número de itens alimentares do questionário num reduzido número de grupos predefinidos (Devlin et al., 2012), representando esses grupos o consumo dietético total tendo em conta as interações entre nutrientes e outros componentes dentro dos mesmos. Como tal, neste estudo foram geradas 26 novas variáveis resultantes da categorização dos 86 itens alimentares do inquérito definida segundo (Lopes et al., 2006), tendo esta como base as similaridades da composição nutricional existentes entre os constituintes em cada grupo (tabela 7.5, Apêndice B).

As variáveis referentes aos valores médios diários de cada nutriente obtidos pela alimentação de cada participante encontram-se descritas na tabela 7.6 do Apêndice B.

3.3 Variáveis referentes às características demográficas dos participantes

Devido ao elevado número de variáveis presentes na base de dados, foi realizada uma seleção inicial das variáveis socioeconómicas, sociodemográficas e respeitantes à dieta e ao estilo de vida dos participantes com interesse para o corrente projeto, que se encontram descritas na tabela 7.7 do Apêndice B. Estas variáveis foram analisadas individualmente no respeitante à ausência de valores para alguma das variáveis medidas para um indivíduo (valores omissos).

Imputação de valores

As variáveis `escolaridade`, `perimetroabd`, `pacotesano`, `fumadorcat`, `horasexercicio`, `suplemento`, `diabetes`, `hipertensao` e `dislipidemia` apresentaram valores omissos. Os valores omissos das variáveis `horasexercicio` (n=4), `suplemento` (n=10), `diabetes` (n=1), `hipertensao` (n=2) e `dislipidemia` (n=1) foram substituídos por 0, uma vez que o facto de serem substituídas por esse valor não alterou o risco de doença de forma significativa (ver codificação das variáveis na tabela 7.7 do Apêndice B).

A variável `fumadorcat` ($n = 7$) foi tratada de modo semelhante ao referido acima, no entanto, foram analisados paralelamente os valores da variável `pacotesano` ($n = 2$), que reporta o número de pacotes de tabaco por ano consumidos pelos participantes. Quando esta variável tomou o valor 0 ou esse valor foi omisso, a variável `fumadorcat` foi substituída por 0 (não fumador). Caso contrário, foi substituída por 1 (fumador), como ocorreu para o participante com ID 58. As únicas 2 omissões da variável `pacotesano` (também omissas para `fumadorcat`) foram igualmente substituídas por 0.

A cada indivíduo omisso para a variável `escolaridade` (n=14), foi atribuído o valor da mediana do grupo de indivíduos com a mesma idade (arredondada à unidade) e sexo.

Por último, foi construído um modelo de regressão linear com base em 962 participantes incluídos no estudo (o participante com ID 2443 foi excluído do modelo) para imputar os valores omissos da variável `perimetroabd` (n=4)

$$E(\texttt{perimetroabd}) = 22.49745 + 9.34948 \times \texttt{sexo} + 0.12567 \times \texttt{idade} + 2.21836 \times \texttt{imc} \quad (3.1)$$

.

A justificação para a eliminação da observação de ID 2443 e diagnósticos dos modelos lineares com e sem essa observação encontram-se descritos no Apêndice C.

Codificação de variáveis

A variável `jafumador` foi derivada pelo agrupamento das categorias fumador (1) e ex-fumador (2) da variável `fumadorcat` e a variável `exercicio` pela dicotomização da variável `horasexercicio`, com o objetivo de simplificar as análises posteriores, como reportado noutros artigos (Islam et al., 2014; Chiu et al., 2014).

Já a variável `esc_cat` foi obtida a partir da categorização da variável `escolaridade`, uma vez que pela observação do histograma referente à distribuição dos participantes segundo os anos de escolaridade (figura 3.1), constatou-se que a maioria dos participantes referiu ter 7 anos de escolaridade. Desse modo a variável original foi repartida em 3 níveis, tendo em consideração os antigos graus de escolaridade. O primeiro nível incluiu o ensino básico e o 2º ciclo atuais (1º ao 6º ano letivo); o segundo nível o 3º ciclo atual (7º ao 9º ano letivo); e o último nível incluiu o ensino secundário atual (10º ao 12º ano letivo) e o ensino superior.

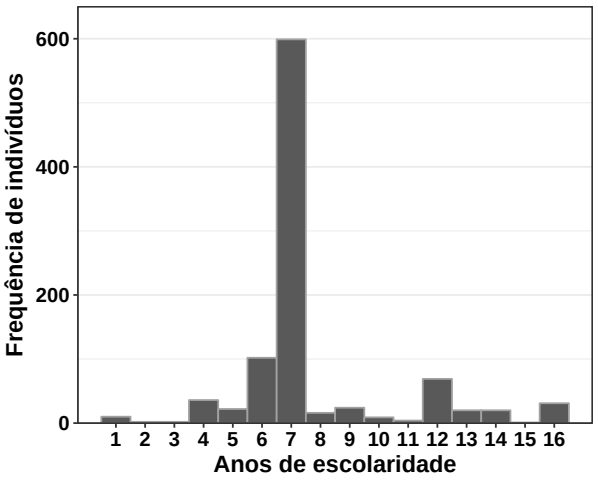


Figura 3.1: Histograma da distribuição dos indivíduos por anos de escolaridade

A variável `suplemento` não foi incluída em futuras análises, dado que a maioria dos indivíduos reportaram consumir suplementos receitados para a osteoporose ou para outras patologias não relacionadas com a visão. Além disso, constatou-se uma falta da especificação dos suplementos ou medicamentos ingeridos e respetiva composição. A variável `desempregado` também não foi incluída nas análises, pois não fornece informação suficiente sobre o estado económico do participante, isto é, se este recebe pensão e de que valor. A `obesidade`, dado estar associada diretamente com o perímetro abdominal e o índice de massa corporal, foi igualmente excluída das subsequentes análises.

Na tabela 3.1 encontra-se a descrição para todas as variáveis (após processamento) candidatas a serem analisadas e posteriormente incluídas no modelo de regressão logística binária.

Tabela 3.1: Descrição das variáveis socioeconómicas, sociodemográficas e respeitantes à dieta e ao estilo de vida dos participantes (após processamento)

	Variável	Tipo de variável	Níveis (codificação)	Descrição
Demografia				
Idade	`idade`	contínua	-	Idade do participante
Sexo	`sexo`	categórica binária	feminino (`0`), masculino (`1`)	Sexo do participante
Escolaridade				
Escolaridade	`esc_cat`	categórica	1-6 anos (`1`), 7-9 anos (`2`), >9 anos (`3`)	Escolaridade do participante
Biometria				
Índice de massa corporal	`imc`	contínua	-	Índice de massa corporal do participante (kg/m^2)
Perímetro abdominal	`perimetroabd`	contínua	-	Perímetro abdominal do participante (cm)
Estilo de vida				
Hábitos tabágicos	`jafumador`	categórica binária	não fumador (`0`), fumador ou ex-fumador (`1`)	Hábitos tabágicos do participante
Pacotes de tabaco por ano	`pacotesano`	contínua	-	Pacotes de tabaco consumidos pelo participante por ano
Atividade física regular	`exercicio`	categórica binária	não (`0`), sim (`1`)	Exercício físico regular realizado pelo participante
Consumo energético total	`cia_energia_kcal`	contínua	-	Consumo energético médio do partipante (kcal/dia)
Consumo de álcool	`alcool`	contínua	-	Consumo de álcool do participante (g/dia)
Comorbilidades				
Diabetes		categórica binária	não (`0`), sim (`1`)	Presença de diabetes no participante
Hipertensão	`hipertensao`	categórica binária	não (`0`), sim (`1`)	Presença de hipertensão no participante
Dislipidemia	`dislipidemia`	categórica binária	não (`0`), sim (`1`)	Presença de dislipidemia no participante

4 Métodos

A estatística multivariada reúne um conjunto de métodos que permitem a análise simultânea de dados caracterizados por duas ou mais variáveis correlacionadas entre si (Reis, 2001). Estes métodos têm-se revelado poderosos na manipulação de dados com muitas variáveis nos diversos ramos da actividade científica.

Na área da epidemiologia nutricional, os métodos estatísticos multivariados que permitem a determinação de padrões alimentares têm-se tornado cada vez mais importantes para a análise da relação entre a dieta e o risco de doenças crónicas (Hu, 2002). Entre os métodos frequentemente utilizados encontram-se a análise em componentes principais e a análise classificatória, que apesar de partilharem o mesmo objetivo, têm diferentes abordagens para atingir o mesmo.

A ACP é frequentemente utilizada para definir padrões alimentares, uma vez que as componentes principais constituem determinadas funções matemáticas das variáveis observadas. Este método utiliza a informação obtida nos questionários de frequência alimentar ou registos dietéticos para identificar dimensões (fatores ou padrões) adjacentes comuns de consumo alimentar. Deste modo, agrega itens alimentares específicos ou grupos alimentares com base no grau de correlação entre esses itens ou grupos (Hu, 2002). Posteriormente, para cada padrão alimentar obtido por ACP é derivado para todos os indivíduos do estudo um *score* sumário, que pode ser utilizado em análise de regressão ou correlação para avaliar a relação entre os vários padrões alimentares e a ocorrência de interesse, neste projeto, a presença ou ausência de DMI (Hu, 2002).

Neste capítulo é feita a descrição das metodologias utilizadas para analisar os dados, de modo a atingir os objetivos a que este projeto se propõe. Primeiramente, serão descritos de forma breve alguns métodos de tratamento de dados a realizar antes da análise estatística (secção 4.1); ao que se segue, na secção 4.2, a descrição da técnica multivariada ACP, e na secção 4.3 da técnica AC, aqui utilizada apenas com o propósito de confirmação e validação dos padrões alimentares obtidos por ACP. Finalmente, na secção 4.4, serão introduzidos os conceitos teóricos básicos da regressão logística binária. Este método será abordado mais detalhadamente, dado que se procurou desenvolver uma estratégia que permitisse simultaneamente a seleção de variáveis e a seleção das formas funcionais das variáveis contínuas, com o objetivo de minimizar o confundimento residual, que poderá conduzir à distorção do verdadeiro efeito dos padrões alimentares no risco da doença.

As análises estatísticas aqui realizadas foram efetuadas com recurso ao *software* **R**, versão 3.3.2, disponível em *Comprehensive R Archive Network* (CRAN) (R Development Core Team, 2016), à exceção

da análise em componentes principais e análise classificatória, que foram efetuadas com recurso ao *software* **Stata** (StataCorp, 2013b), versão 13.0, adquirido e disponibilizado pela AIBILI no âmbito do corrente projeto. Em anexo, encontra-se uma tabela que resume as principais bibliotecas e funções usadas nesta análise (tabela 7.17, Apêndice D).

4.1 Pré-tratamento de dados

Antes da realização da análise estatística é geralmente realizado um pré-processamento dos dados. A existência de observações omissas e de pseudo-variáveis (variáveis de valor constante) é relativamente comum na maioria dos estudos e pode ter um impacto significativo nas conclusões retiradas dos dados. Vários problemas poderão surgir com a existência de valores omissos, como a redução do poder estatístico, viés na estimação de parâmetros, redução da representatividade das amostras e complicação das análises estatísticas do estudo. Tal exige a preparação dos dados de modo a que estas situações não ocorram e que a matriz de dados seja completa. Neste projeto procedeu-se à imputação dos valores omissos (ver secção 3.3) segundo várias metodologias (Kalton and Kasprzyk, 1986), que são:

- **Imputação dedutiva:** o valor imputado é deduzido com base na lógica, sem um mecanismo probabilístico, a partir de informação conhecida, nomeadamente em inquéritos anteriores que utilizam as mesmas questões e amostra;
- **Imputação pela mediana da classe:** a amostra total é dividida em classes de acordo com a combinação dos valores das covariáveis auxiliares a serem usadas na imputação. Dentro de cada classe de imputação, a correspondente mediana da classe é atribuída ao(s) valor(es) omisso(s);
- **Imputação via análise de regressão linear múltipla:** é construído um modelo de regressão onde a variável dependente é aquela que apresenta o(s) valor(es) omisso(s) a imputar e as variáveis independentes são as covariáveis consideradas importantes para a explicação da primeira. A equação resultante é utilizada para estimar o(s) valor(es) omisso(s) da variável de interesse, onde o valor predito da regressão é atribuído a esse(s) valor(es).

Adicionalmente, pode também ser necessária a centralização ou a normalização dos dados. Definam-se primeiramente as seguintes matrizes:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}; \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}; \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{n2} & \dots & \sigma_p^2 \end{bmatrix},$$

onde $\mathbf{X}$ representa a matriz de dados, de n observações, do vetor aleatório $\mathbf{X} = (X_1, X_2, \cdots, X_p)$; $\boldsymbol{\mu}$ o vetor de valores médios de $\mathbf{X}$; e $\boldsymbol{\Sigma}$ a respetiva matriz de covariâncias, onde $\sigma_{ij} = Cov(X_i, X_j)$ com $i = 1, 2, \cdots, p$ e $j = 1, 2, \cdots, p$.

A estandardização das variáveis aleatórias obtida por subtração do valor médio e divisão pelo desvio padrão,

$$\mathbf{Z}_j = \frac{\mathbf{X}_j - \mu_j}{\sigma_j}, \tag{4.1}$$

com $j = 1, 2, \cdots, p$, é necessária quando as variáveis se encontram medidas em diferentes escalas ou unidades ou são de natureza distinta, de modo a que todas as variáveis passem a ter variância

unitária. Note-se que, após a estandardização, a influência das variáveis de pequena variância tende a ser inflacionada, enquanto que a influência das variáveis de grande variância tende a ser reduzida.

A matriz de covariâncias das variáveis estandardizadas será então igual à matriz de correlações do conjunto das variáveis iniciais, definida por:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix},$$

onde $\rho_{ij} = \frac{Cov(X_i,X_j)}{\sigma_i \sigma_j} = Corr(X_i, X_j)$ com $i = 1, 2, \cdots, p$ e $j = 1, 2, \cdots, p$.

Quando as matrizes $\boldsymbol{\Sigma}$ e $\boldsymbol{\rho}$ não são conhecidas, como acontece neste estudo, são substituídas pelas matrizes amostrais **S** e **R**, respetivamentes, que correspondem às suas estimativas. Neste projeto, as variáveis encontram-se medidas na sua maioria em gramas, no entanto, algumas estão expressas em mililitros pelo que se irão utilizar os dados estandardizados.

4.2 Análise em componentes principais

A análise em componentes principais (ACP) é um método de análise de dados multivariados introduzido pelo estatístico Karl Pearson em 1901 e, mais tarde, desenvolvido por Hotelling em 1933 (Bibby et al., 1980). Um dos objetivos principais deste método é a redução da dimensionalidade dos dados, preservando o máximo de informação relevante possível, através da análise das inter-relações entre um grande número de variáveis quantitativas. Esta técnica permite a simplificação de grandes conjuntos de dados e, consequentemente, uma melhor compreensão e interpretação da estrutura inter-relacional inerente às variáveis originais (Paul and Al Suman, 2013).

Esta ferramenta estatística baseia-se na aplicação de uma transformação linear ao conjunto das variáveis originais correlacionadas, com vista à obtenção de um novo conjunto de variáveis não correlacionadas, as componentes principais.

As componentes principais são combinações lineares das variáveis originais e estimadas com o propósito de reter o máximo de informação possível, em termos de variação total associada aos dados iniciais. A quantidade de informação que uma componente principal consegue explicar é medida pela respetiva variância, que mede o grau de diferenciação entre elementos de um conjunto de dados. Deste modo, as componentes são estimadas por ordem de importância, sendo a primeira componente estimada a mais importante, pois descreve uma maior variabilidade dos dados, seguida pela segunda componente, e assim por diante, até à última componente estimada.

Com este método pretende-se reter o menor número de componentes principais, ou seja, substituir as p variáveis originais por r (com $r << p$) componentes principais, com perda mínima de informação. A existência de possível redundância das variáveis originais é assim eliminada pela aplicação deste método, o que permite reduzir a dimensionalidade dos dados, pois a informação fornecida pelas variáveis originais pode ser explicada com um menor número variáveis, as componentes principais. Note-se que caso as variáveis originais estejam fracamente correlacionadas, as componentes principais

vão, em grande parte, coincidir com as variáveis originais, sendo a redução pouco significativa.

O modelo matemático

As componentes principais são representadas segundo o modelo matemático

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = \mathbf{a}_j'\mathbf{X}, \tag{4.2}$$

onde $X_1, \cdots, X_p$ são as variáveis originais; Y_j, com $j = 1, \cdots, r$ ($r \leq p$), é a j-ésima componente principal obtida; $a_{1j}, \cdots, a_{pj}$ são os coeficientes da combinação linear, representando o peso da i-ésima variável ($i = 1, \cdots, p$) na j-ésima componente principal.

Na determinação dos coeficientes das componentes principais, têm de ser satisfeitas as seguintes condições:

1. $Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_r)$, com $r \leq p$;
2. Quaisquer duas componentes principais são não correlacionadas: $Corr(Y_i, Y_j) = 0, \forall_{i,j}$;
3. Em qualquer componente principal a soma dos quadrados dos coeficientes que engloba é 1 (para $Y_j : a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2 = 1 \Leftrightarrow \mathbf{a}_j'\mathbf{a}_j = 1$), isto é, os vetores $\mathbf{a}_j$ são normados.

Neste trabalho, as componentes principais foram estimadas no *software* **Stata**, com recurso ao comando **pca**, que utiliza o método algébrico de decomposição espetral da matriz de correlações $\boldsymbol{\rho}$, ou de covariâncias $\boldsymbol{\Sigma}$, de modo a calcular os coeficientes das componentes principais. Este método será seguidamente descrito.

Seja $\mathbf{C}$ a matriz $(p \times p)$ de variância-covariância ou de correlação e considere-se ainda os valores próprios de $\mathbf{C}$ e os vetores próprios normalizados. Tem-se $\mathbf{CQ} = \mathbf{Q\Lambda}$, com $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, onde $\mathbf{\Lambda}$ é a matriz diagonal dos valores próprios,

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix},$$

e $\mathbf{Q}$ a matriz ortogonal dos vetores próprios. Como $\mathbf{Q}^\top = \mathbf{Q}^{-1}$, então $\mathbf{C} = \mathbf{Q\Lambda Q}^\top$.

Para a determinação das componentes principais, começa-se portanto por calcular os valores próprios da matriz $\mathbf{C}$ $(\lambda_1, \lambda_2, \cdots, \lambda_p)$, para depois se determinarem os vetores próprios associados aos diferentes valores próprios, que correspondem aos coeficientes das componentes principais $(\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_p)$. A componente principal j é, afinal, o vetor próprio associado ao valor próprio j, onde $Var(Y_j) = \lambda_j$, $j = 1, 2, \cdots, p$.

Das condições acima apresentadas retiramos que a primeira componente principal extraída (Y_1), é a componente com maior variância; Y_2 é a componente com a segunda maior variância, sujeita à condição de ser não correlacionada com Y_1; Y_3 é a componente com a terceira maior variância, sujeita à condição de ser não correlacionada com as componentes anteriores; e assim sucessivamente até à obtenção da última componente, Y_p. Isto é, a ordenação das componentes principais é feita através da ordenação dos valores próprios $(\lambda_1 > \lambda_2 > \cdots > \lambda_p \geq 0)$. É de salientar que as componentes obtidas são diferentes quando se usa a matriz de covariâncias e quando se usa a matriz de correlações, dado não

existir invariância de escala na ACP (Gnanadesikan, 2011).

Uma vez obtidas as componentes principais, os seus valores númericos (ou *scores*) podem ser calculados para cada indivíduo da amostra:

$$y_{ij} = a_{1j}x_{i1} + a_{2j}x_{i2} + \cdots + a_{pj}x_{ip}, \tag{4.3}$$

onde y_{ij} representa o *score* do i-ésimo indivíduo ($i = 1, \cdots, n$) para a j-ésima componente principal ($j = 1, 2, \cdots, r$, em que r representa o número de componentes retidas).

Aplicando esta transformação aos dados, obtém-se uma nova matriz de dados, a matriz de *scores* (**Y**), com dimensão $(n \times r)$:

$$\mathbf{Y} = \mathbf{X}.\mathbf{A} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1r} \\ y_{21} & y_{22} & \cdots & y_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nr} \end{bmatrix},$$

onde **A** é a matriz cujas colunas são constituídas pelos p vetores próprios.

No contexto deste projeto, a ACP permite, assim, reduzir a dimensionalidade dos dados, onde se passa a trabalhar com um menor número de variáveis (componentes principais retidas). Posteriormente, uma vez que as novas variáveis não são correlacionadas entre si (independentes), essas poderão ser incluídas num modelo de análise de regressão logística binária com fim a estudar a respetiva associação com a doença.

Número de componentes a reter

Um dos grandes desafios na análise de componentes principais é a decisão do número de componentes principais a reter. Apenas as primeiras componentes são utilizadas, pois são as mais importantes na explicação da variação associadas às variáveis originais (Jolliffe, 2002).

Vários critérios se encontram na literatura (Fransen et al., 2014; Moeller et al., 2007), entre eles:

- **Percentagem de variância total explicada**: um dos primeiros critérios frequentemente utilizado é reter o número de componentes principais que expliquem pelo menos 80% da variabilidade total dos dados. Sendo λ_j a variância da j-ésima componente principal e $\sum_{l=1}^{p} \lambda_l$ a variância total, tem-se que $\frac{\lambda_j}{\sum_{l=1}^{p} \lambda_l} \times 100$ é a percentagem de variância total explicada pela j-ésima componente principal e $\frac{\sum_{j=1}^{k} \lambda_k}{\sum_{l=1}^{p} \lambda_l} \times 100$ a percentagem de variância total explicada pelas primeiras k componentes principais. Este valor deverá então ser igual ou superior ao limite de variabilidade total decidido pelo utilizador. Neste projeto este critério não foi utilizado uma vez que depende grandemente do número de total de variáveis incluídas na análise (Chiu et al., 2014).
- **Critério da média ou de Kaiser**: esta regra consiste em reter apenas as componentes principais às quais corresponde valores próprios acima da média:

$$\bar{\lambda} = \frac{1}{p} \sum_{j=1}^{p} \lambda_j. \tag{4.4}$$

Note-se que no caso particular de os dados estarem centrados e estandardizados, a média será 1. No gráfico *scree plot* apresentado abaixo como exemplo (figura 4.1), podemos facilmente verificar que apenas 4 componentes principais obedecem a este critério, sendo aquelas a ser retidas.

– **Análise gráfica**: utilizar um gráfico *scree plot* (figura 4.1), proposto por Cattell em 1966, onde se representam os pontos de abcissa j e ordenada igual ao j-ésimo valor próprio ou à percentagem de variância explicada pela j-ésima componente principal ($j = 1, 2, \cdots, p$). Através da análise do gráfico, devem ser retidas as componentes principais que contribuem com maior quantidade de informação e que se destacam de modo acentuado das restantes. No caso concreto do gráfico 4.1, o número de componentes a reter seria 3 ou 4, uma vez terem valores próprios mais destacados dos restantes.

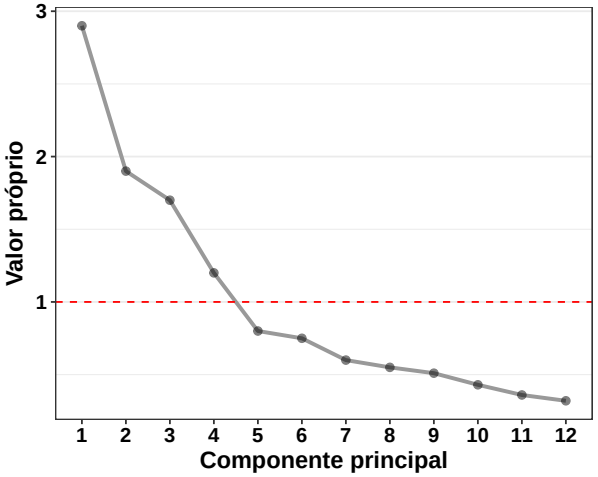


Figura 4.1: Exemplo da representação gráfica dos valores próprios em função do número de componentes principais (gráfico *scree plot*). A linha tracejada no valor próprio 1 ajuda a visualizar quais as componentes principais que obedecem ao critério de Kaiser (variáveis originais centradas e estandardizadas).

Interpretação das componentes principais

Embora nem sempre seja possível, é vantajoso atribuir um significado a cada uma das componentes principais, permitindo uma melhor interpretação e utilização dos resultados da análise. A interpretação das componentes principais pode ser feita utilizando os coeficientes das combinações lineares (a_{ij}) e as correlações entre as variáveis iniciais e as componentes principais, denominados *loadings* (l_{ij}).

Os coeficientes das componentes principais (a_{ij}) representam a importância relativa da variável i na componente principal j. Já os *loadings*, obtidos por

$$l_{ij} = a_{ij}\sqrt{\lambda_j}, \tag{4.5}$$

com $i = 1, 2, \cdots, p$ e $j = 1, 2, \cdots, r$; representam a correlação entre as variáveis originais e as componentes principais, fornecendo a indicação de como as variáveis originais são importantes para a formação das componentes principais.

Existem várias regras que ajudam na decisão de quais as variáveis a serem usadas na interpretação de uma componente principal, sendo estas baseadas no coeficiente de correlação entre a i-ésima variável (X_i) e a j-ésima componente principal (Y_j):

$$Corr(X_i, Y_j) = \frac{Cov(X_i, Y_j)}{\sqrt{Var(X_i)}\sqrt{Var(Y_j)}} = \frac{a_{ij}\lambda_j}{\sigma_i\sqrt{\lambda_j}} = \frac{a_{ij}\sqrt{\lambda_j}}{\sigma_i} = \frac{l_{ij}}{\sigma_i}. \tag{4.6}$$

Valores de $Corr(X_i, Y_j)$ próximos de 1 ocorrem quando a variável é importante para a formação da componente, tomando valores próximos de 0 caso não seja importante.

Neste projeto, apenas os grupos de alimentos que apresentaram um *loadings* com valor absoluto igual ou superior a 0.2 ($|l_{ij}| \geq 0.2$) foram tidos em conta na interpretação e definição dos padrões alimentares; este limiar é frequentemente reportado na literatura (Schulze et al., 2003).

Para uma melhor interpretação das r componentes principais retidas, vários autores recomendam a aplicação de uma rotação dos eixos. As rotações podem ser ortogonais ou oblíquas, no entanto, as ortogonais são as mais utilizadas, destacando-se a rotação *varimax*, *quartimax* e *equimax*. Neste projeto, foi apenas utilizada a rotação *varimax*, a opção *default* no **Stata**, portanto, será a única doravante mencionada (StataCorp, 2013a).

A rotação *varimax* consiste na transformação da solução inicial através da multiplicação de uma matriz de rotação ortogonal pela matriz dos *loadings*, com o objetivo de maximizar a soma das variâncias dos quadrados dos *loadings* de cada coluna da matriz $\mathbf{\Lambda} = [\lambda_{ij}]$. Deste modo, são aumentados os valores absolutos iniciais dos *loadings* de elevada magnitude e são diminuídos os de baixa magnitude, havendo maior distinção entre os *loadings* significantes e os insignificantes (Kaiser, 1958).

Este é um procedimento que, apesar de se encontrar frequentemente publicado (Abdi, 2003; Islam et al., 2014; Moeller et al., 2007), é bastante controverso entre os cientistas, pois altera algumas das propriedades das componentes principais. Neste projeto foi realizada ACP sem e com rotação *varimax* e comparados os resultados obtidos por cada uma das abordagens, de modo a escolher aquela que gerasse uma melhor interpretabilidade das componentes principais.

4.3 Análise classificatória

A análise classificatória (AC) consiste noutra abordagem de análise de dados quantitativos multivariados, sendo utilizada para agrupar objetos (casos ou observações) em grupos relativamente homogéneos denominados *clusters*. Quando não existe informação prévia sobre os grupos, este método classifica-se como sendo não supervisionado, pois a formação dos *clusters* é sugerida unicamente pelos dados (Landau and Ster, 2010).

Esta ferramenta estatística permite reduzir a dimensionalidade do conjunto de dados pela sua classificação dos indivíduos em grupos atendendo à semelhança das suas propriedades inerentes, ao contrário de outras técnicas multivariadas (análise discriminante, análise de componentes principais, entre outras), que analisam as relações existentes entre as variáveis originais (Yim and Ramdeen, 2015).

Os objetos dos *clusters* construídos com base nesta técnica tendem a ter, portanto, propriedades semelhantes entre si, mas diferentes de objetos não pertencentes ao mesmo *cluster*, sendo os *clusters* mutuamente exclusivos e exaustivos (Devlin et al., 2012).

Existe uma variedade de diferentes processos de *clustering* para formar os grupos de objetos, cada um requerendo diferentes decisões a tomar por parte do estatístico antes de ser efetuada a análise classificatória, pelo que a escolha do processo a utilizar é crucial (Yim and Ramdeen, 2015).

Neste projeto serão discutidos apenas os dois tipos principais de métodos de AC utilizados na área da epidemiologia nutricional: os métodos hierárquicos e os métodos de partição (ou não hierárquicos). Cada um destes métodos segue uma abordagem diferente para agrupar os objetos mais similares num *cluster* e para determinar a que *cluster* cada objeto pertence (Mooi and Sarstedt, 2010).

4.3.1 Classificação hierárquica

Os processos de classificação hierárquica são caracterizados pela obtenção de uma estrutura em árvore (dendograma) estabelecida no curso da análise (ver figura 4.2 para exemplo de dendograma).

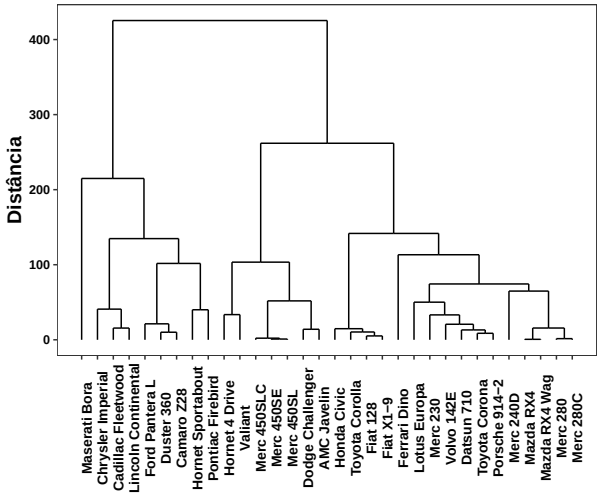


Figura 4.2: Exemplo de dendograma. Base de dados utilizada proveniente da biblioteca **mtcars** do *software***R**

Estes processos dividem-se em duas categorias distintas na formação de *clusters*: métodos hierárquicos aglomerativos ou divisivos.

Nos métodos aglomerativos, os grupos são formados segundo algoritmos que se baseiam em sucessivas fusões dos objetos em novas classes, terminando numa só; inversamente, nos métodos divisivos, têm lugar sucessivas partições de uma só classe que abrange todos os objetos até se obterem tantas classes como objetos (Yim and Ramdeen, 2015).

Em termos práticos, os métodos aglomerativos são os mais utilizados e os únicos disponíveis no *software* **Stata**, dado o tempo de computação particularmente elevado subjacente aos métodos divisivos (StataCorp, 2013a). Por esta razão, neste projeto serão apenas abordados os métodos aglomerativos.

Existem diversos tipos de processos aglomerativos que poderão ser aplicados. No entanto, para decidir qual o mais apropriado consoante a natureza dos dados disponíveis e o(s) objetivo(s) a atingir, é indispensável o conhecimento do modo como as similaridades e dissimilaridades são medidas entre os pares de objetos (Mooi and Sarstedt, 2010).

Seleção da medida de distância

No presente contexto, existem dois tipos principais de medidas usadas para estimar a relação entre objetos: medidas de distância e medidas de similaridade. Ambas as medidas são frequentemente definidas como o oposto uma da outra, ou seja, quando a distância entre dois casos aumenta, a sua similaridade diminui. No entanto, uma importante distinção deverá ser feita: enquanto que ambas as medidas refletem o padrão dos *scores* das variáveis selecionadas, apenas a medida de distância tem em conta a elevação desses *scores* (Clatworthy et al., 2005). Por esta razão, serão apenas descritas e utilizadas neste projeto as medidas de distância.

Considere-se novamente uma matriz de dados $\mathbf{X}$ de ordem $n \times p$, que representa o conjunto de n observações e p variáveis, e a respetiva matriz de covariâncias $\mathbf{\Sigma}$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}; \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{n2} & \dots & \sigma_p^2 \end{bmatrix},$$

onde $\sigma_{ij} = Cov(X_i, X_j)$ com $i = 1, 2, \cdots, p$ e $j = 1, 2, \cdots, p$. Caso a matriz $\mathbf{\Sigma}$ não seja conhecida, como acontece neste estudo, é substituída pela sua estimativa, a matriz de covariâncias amostral ou empírica $\mathbf{S}$.

A distância entre dois objetos $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$ e $\boldsymbol{x}_j = (x_{j1}, \cdots, x_{jp})$ ($i = 1, 2, \cdots, n$ e $j = 1, 2, \cdots, n$) será denotada por d_{ij}.

Uma medida de distância (ou métrica) caracteriza-se pelas seguintes propriedades (Lovric, 2011):

1. $d_{ij} \geq 0, \forall_{i,j}$ (não-negatividade),
2. $d_{ii} = 0, \forall_i$ (nula apenas para pontos coincidentes),
3. $d_{ij} = \delta_{ji}, \forall_{i,j}$ (simetria),
4. $d_{ij} \leq d_{im} + d_{mj}, \forall_m \in \boldsymbol{X}$ (desigualdade triangular).

Várias estatísticas poderão ser utilizadas como medidas de distância na AC. A escolha depende principalmente da natureza das variáveis, ou seja, se são contínuas ou categóricas (Yim and Ramdeen, 2015).

Quando as variáveis em questão são contínuas, como acontece no estudo decorrente, a medida mais frequentemente empregue é a distância euclidiana e as distâncias que derivam desta, como o quadrado da distância euclidiana e a distância de Mahalanobis (Everitt et al., 2011; Reis, 2001).

A distância euclidiana entre quaisquer dois pontos num espaço p-dimensional é dada por

$$d_{i,j} = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2}. \tag{4.7}$$

Apesar de esta distância ser muito utilizada, em determinadas situações, como quando as variáveis em consideração estão medidas em diferentes escalas ou níveis, ou quando as variáveis têm variâncias muito

diferentes ou verificam ser muito correlacionadas, ou quando existem dados omissos, é preferível utilizar a distância de Mahalanobis. Esta distância foi a utilizada neste trabalho, uma vez que resolve não só o problema das escalas, como também o problema dos efeitos das correlações entre as variáveis:

$$d_{i,j} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \mathbf{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}, \tag{4.8}$$

onde **S** representa a matriz empírica de covariâncias da matriz de dados **X**. Note-se que as diferenças entre as variáveis i e j são mais valorizadas quando são pouco correlacionadas e menos valorizadas quando têm maior correlação. Deste modo, a distância de Mahalanobis corrige para a estrutura de correlação das variáveis originais, compensando pela colinearidade entre essas variáveis.

Seleção do algoritmo aglomerativo

Outro critério a ser tido em conta neste tipo de análise é a estatística a ser usada no cálculo da distância ou similaridade entre os *clusters* (medida de ligação). Esta estatística está na base dos vários processos aglomerativos existentes, que são distinguidos pela maneira como definem a distância entre um *cluster* novo e um objeto, ou entre outros *clusters* na solução (Mooi and Sarstedt, 2010).

É importante salientar que cada um destes algoritmos pode gerar resultados completamente diferentes quando usados no mesmo conjunto de dados, uma vez que cada um tem propriedades específicas. Entre os principais métodos encontram-se o da ligação simples, da ligação completa, da ligação média e de Ward.

No âmbito da epidemiologia nutricional, o método de Ward é o método hierárquico aglomerativo que se encontra mais reportado na literatura. Este método de agrupamente foi proposto por Ward em 1963, baseando-se na alteração da variação dentro e entre os grupos a serem formados a cada passo do processo do agrupamento. É também conhecido como método da variância mínima pois,ao contrário dos outros métodos aglomerativos, tem como objetivo combinar os objetos cuja fusão resulte no menor aumento da variância total dentro dos *clusters* (Shalizi, 2009). Para o método de Ward, a distância entre dois *clusters*, C_A e C_B, é, portanto, a medida do incremento ($I_{C_A C_B}$) que a soma de quadrados das distâncias aos respetivos centróides sofre quando esses *clusters* se fundem

$$I_{C_A C_B} = \frac{n_A n_B}{n_A + n_B}\, d^2(\bar{x}_A, \bar{x}_B), \tag{4.9}$$

onde $d^2(\bar{x}_A, \bar{x}_B)$ representa o quadrado da distância euclidiana entre os vetores médios dos *clusters* C_A e C_B, e n_A e n_B são os números de observações nos *clusters* C_A e C_B, respetivamente.

De maneira geral, num algoritmo do tipo aglomerativo é primeiramente determinada a matriz inicial de distâncias **D** para os n objectos:

$$\mathbf{D} = \begin{bmatrix} 0 & \dots & \dots & \\ d_{21} & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{bmatrix}.$$

O cálculo da matriz de distâncias é então seguida de vários ciclos que se baseiam em dois passos repetidos até se obter apenas um grupo ou *cluster*:

1. Determinação do menor elemento da matriz **D** e junção das respetivas classes numa só;

2. Cálculo dos valores de distância entre a nova classe e todas as restantes e atualização da matriz **D**.

4.3.2 Classificação não hierárquica

A classificação não hierárquica assenta em diferentes princípios relativamente à classificação hierárquica e os seus resultados não constituem hierarquias, tal como o nome sugere (Quintal, 2006).

Algumas das vantagens sobre os métodos hierárquicos consistem no facto de os métodos não hierárquicos, em cada passo do algoritmo, serem capazes de reagrupar os objectos em *clusters* diferentes daqueles em que foram colocados anteriormente e não necessitarem de calcular e armazenar uma nova matriz de distâncias, podendo ser aplicados a matrizes de dados muito grandes. No entanto, nestes métodos é necessário definir *a priori* o número final de grupos, sendo muitas vezes difícil, dado o desconhecimento da estrutura dos dados. A opção aqui utilizada foi a aplicação prévia do método hierárquico de Ward aos dados para ajudar a determinar o número de *clusters* a reter da amostra (Thorpe et al., 2016).

Os métodos não hierárquicos podem ser métodos de partição, métodos baseados em modelos, métodos difusos ou métodos de sobreposição. Estes diferem essencialmente na forma como se realiza a primeira agregação dos objectos em *clusters*, e no modo como as distâncias entre os centróides dos *clusters* e os objectos são medidas. No âmbito do corrente projeto serão apenas focados os métodos de partição.

Os métodos de partição, à semelhança dos métodos hierárquicos, também dispõem de uma grande variedade de algoritmos a serem aplicados. Entre estes, na área da epidemiologia nutricional, o método *K-means* é o mais reportado na literatura (Moeller et al., 2007). Como tal, foi o método adoptado neste estudo, assim como uma variante sua, o método *K-medians*, que serão de seguida descritos.

Método *K-means* e *K-medians*

Os métodos *K-means* e *K-medians* aplicam-se apenas a objectos, não a variáveis, operando sobre uma matriz de dados inicial **X**, e não sobre uma matriz de distâncias **D** (como se verifica nos métodos hierárquicos). Nestes métodos é construída uma partição a partir dos dados, ou seja, uma coleção de grupos disjuntos de objetos cuja reunião constitui o conjunto de objetos inicial.

Estes algoritmos processam-se de modo a particionar os dados através da minimização da variação dentro de cada *cluster*, sendo esta utilizada como uma medida para formar *clusters* homogéneos. Consequentemente, não é necessária a escolha da medida de distância previamente à análise, como acontece nos métodos hierárquicos (Mooi and Sarstedt, 2010).

De maneira geral, os procedimentos dos métodos de partição baseiam-se em vários passos (Johnson et al., 2002). Primeiramente é feita a seleção de uma partição inicial dos objectos em k *clusters*, que neste trabalho foi realizada de modo aleatório; seguida do cálculo dos centróides (neste caso, centro geométrico dos *clusters*) para cada um dos k *clusters* e cálculo das distâncias euclidianas dos centros dos *clusters* a cada um dos objetos; cada objeto é então agrupado ao *cluster* de cujo centróide se encontra mais próximo (StataCorp, 2013a).

Deste modo, obtém-se uma partição inicial dos dados nos vários *clusters* da solução. Novamente, são calculados os centróides dos *clusters* relocalizados; são calculadas as distâncias dos centróides a cada objeto; e são atribuídos os objetos ao *cluster* de centro mais próximo. Note-se que, dada a nova localização do centro dos *clusters*, poder-se-á obter uma solução diferente.

Estes passos são repetidos até não ocorrer variação significativa na distância mínima de cada objecto da matriz de dados a cada um dos centróides dos k *clusters*, ou até que o número máximo de iterações ou o critério de convergência (ou seja, não exista mudança na composição dos *clusters*), definido pelo utilizador, seja alcançado. O **Stata** especifica 10000 iterações como número máximo (*default*).

No método *K-means* os centróides são calculados através da computação do cálculo dos vetores médios dos objetos incluídos no *cluster*, tendo em conta cada uma das variáveis. Este método apresenta várias limitações além da escolha prévia do número de partições, apontando-se entre elas a sensibilidade a *outliers*. Por esta razão, neste trabalho foi também implementado o método *K-medians*, uma variante do método *K-means*, onde em vez de se calcular a média para cada *cluster* para determinar o respetivo centróide, é calculada a mediana dos objetos de cada *cluster*. Note-se que pelo facto da mediana ser menos sensível que a média perante a presença de *outliers*, este algoritmo é, portanto, uma alternativa mais resistente (Rai, 2011).

4.3.3 Número de *clusters* a reter

Por fim, a escolha do número de *clusters* a reter dos dados é a última decisão a tomar na análise classificatória. Neste estudo foram seguidos vários passos para determinar o número de *clusters*, de acordo com o estudo *A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians* (Thorpe et al., 2016).

Primeiramente, como mencionado nas secções acima, foi aplicada classificação hierárquica segundo o método de Ward. Nos métodos hierárquicos a análise gráfica do respetivo dendograma é uma grande ajuda na decisão do número de *clusters* a utilizar, sendo que se pode optar por parar o processo quando a distância entre classes obtida em ciclos sucessivos excede um determinado valor, ou quando as diferenças entre distâncias sofrem um aumento abrupto.

O *software* **Stata** dispõe também de critérios de paragem que apoiam o utilizador nesta decisão, nomeadamente o índice pseudo-F de Calinski e Harabasz (1974) e o índice Je(2)/Je(1) Duda-Hart (1973), apontados como sendo os dois melhores (Milligan and Cooper, 1985). Em ambas as regras, valores elevados são indicadores de *clustering* mais distinto.

No presente estudo, optou-se pela utilização da regra de Calinski-Harabasz, que pode também ser aplicada com os métodos não hierárquicos, ao contrário do índice Je(2)/Je(1) Duda-Hart (ver mais detalhes em (StataCorp, 2013a)).

Após consideração dos resultados obtidos pelo método Ward, foram aplicados os métodos *K-means* e *K-medians*, com as respetivas soluções a variar de 2 a 8 *clusters* (Newby et al., 2004). Foi então aplicada a regra de paragem de Calinski-Harabasz a cada uma das soluções, de modo a determinar qual a solução para cada método com *clustering* mais distinto (Duda and Hart, 1973; Caliński and Harabasz, 1974).

Finalmente, a interpretabilidade das soluções obtidas para cada um dos métodos não hierárquicos foi analisada, de modo a confirmar a solução final.

4.3.4 Validação

Dada a sensibilidade da análise classificatória, foi realizada a avaliação da estabilidade e validade das soluções obtidas pelo método *K-means* e *K-medians*, de modo a decidir qual o método e a solução final a utilizar (Thorpe et al., 2016).

Primeiramente, os *clusters* de cada solução foram numerados, e foi-lhes atribuída uma designação provisória, consoante os grupos alimentares com uma média significativamente alta ou reduzida de consumo. De seguida, a amostra total foi dividida em duas metades aleatórias e uma dessas subamostras foi escolhida aleatoriamente para uma nova aplicação do mesmo processo de análise classificatória aplicado na solução a validar. A concordância entre os *clusters* da amostra total e dos *clusters* da subamostra aleatória foi então medida através da estatística Kappa de Cohen (κ), abaixo descrita. Este processo de divisão aleatória e obtenção da medida de concordância e associação foi realizado 10 vezes, de modo a reduzir o impacto da seleção aleatória da subamostra em que a análise é aplicada (Siou et al., 2011), e foram calculados os respetivos valores médios para cada solução, tendo a solução com maior valor médio da estatística Kappa sido considerada a solução final por apresentar uma maior estabilidade e reprodutibilidade.

A estatística κ mede a concordância entre dois classificadores, que neste caso são a amostra original e a metade aleatória, que classificam N itens em C categorias exclusivas:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \tag{4.10}$$

onde p_o é a concordância relativa observada entre classificadores, e p_e é a probabilidade hipotética de concordância aleatória, usando os dados observados para calcular as probabilidades de cada observador atribuir aleatoriamente a cada uma das categorias. O valor da estatística Kappa geralmente varia entre 0 e 1, podendo tomar valores negativos se a probabilidade de concordância observada for menor que a esperada. Quanto maior a estatística, maior a concordância, sendo que o valor 1 indica concordância completa.

É importante também salientar que durante a realização do procedimento acima descrito, foi atribuída uma numeração pelo *software* aos *clusters* de forma aleatória. Como tal, esta teve de ser posteriormente corrigida, de modo a coincidir com a utilizada na respetiva amostra original, com vista à obtenção de medidas de concordância adequadas. O processo de correção e atribuição da numeração a cada *cluster* foi baseado análise conjunta dos resultados de três procedimentos abaixo descritos.

O primeiro procedimento consiste na análise dos valores médios do consumo nos vários grupos alimentares, de forma a verificar quais obtinham um valor significativamente elevado ou reduzido em determinado *cluster*, e assim a associar a cada *cluster* da amostra total o *cluster* da subamostra que fosse mais semelhante na sua estrutura. No entanto, muitas vezes esta análise não permitiu retirar conclusões claras, tanto porque os *clusters* da subamostra apresentavam uma composição semelhante com vários *clusters* da amostra original, ou porque alguns *clusters* da subamostra em nada se assemelhavam com aqueles.

Como tal, outro procedimento foi tido em conta para ajudar nesta decisão. Neste, as médias dos valores de cada grupo alimentar j ($j = 1, 2, \cdots, 26$) foram calculadas para um determinado *cluster* i da amostra original ($i = 1, 2, \cdots, k$, onde k representa o número total de *clusters* da solução a validar). Denote-se esta média por $\bar{x}_{ij}$. Do mesmo modo, calcularam-se as médias em cada *cluster* l ($l = 1, 2, \cdots, k$) da subamostra aleatória para cada grupo alimentar ($\bar{y}_{lj}$). Para cada *cluster* i da amostra total foi efetuado, assim, o seguinte cálculo para todos os *clusters* l da subamostra aleatória:

$$\sum_{j=1}^{26} (\bar{x}_{ij} - \bar{y}_{lj})^2, \tag{4.11}$$

sendo atribuída ao *cluster* da subamostra que apresente o menor valor desta estatística a numeração adjacente ao *cluster* da amostra original.

O último procedimento apenas foi utilizado quando a aplicação do primeiro e do segundo procedimento gerou dúvidas relativamente à atribuição da numeração entre dois *clusters* da subamostra aleatória. Este consistiu na análise das proporções de indivíduos nos *clusters* da amostra original e das proporções relativas aos *clusters* da subamostra, sendo que a numeração relativa a determinado *cluster* da amostra original foi atribuída ao *cluster* da subamostra com a respetiva proporção mais próxima da desse *cluster*.

4.4 Regressão logística binária

A análise de regressão é uma das ferramentas estatísticas mais importantes na área da epidemiologia, uma vez que permite investigar o efeito de apenas uma ou várias variáveis explicativas ou covariáveis (como a exposição a determinadas condições, características dos indivíduos, fatores de risco) numa variável resposta, como o desenvolvimento de uma doença ou sintoma, mortalidade, entre outras. Frequentemente, em investigação clínica, essa variável resposta caracteriza-se por ser binária resultante, por exemplo, da presença ou ausência de doença, como é o caso dos dados analisados neste projeto.

A regressão logística binária é o método mais utilizado quando a variável resposta é desta natureza, estando inserida nos denominados modelos lineares generalizados (MLG), apresentados pela primeira vez em 1972 por Nelder e Wedderburn e que consistem numa extensão dos modelos lineares (Nelder and Baker, 1972). A análise de regressão logística binária permite o uso de um modelo de regressão para se estimar a probabilidade de um evento específico, avaliando a influência das variáveis explicativas quanto à variável resposta, assim como os efeitos das suas potenciais interações. É também importante salientar que, com esta metodologia, as variáveis explicativas podem ser categóricas ou quantitativas (Bender, 2009).

A característica mais atrativa de um modelo de regressão logística é tanto não assumir a linearidade na relação entre as covariáveis e a variável resposta, como também não requerer a normalidade das distribuições das variáveis. A homocedasticidade também não é assumida e, no geral, tem requisitos menos rigorosos que os dos modelos de regressão linear (Wright, 1995). Deste modo, a regressão logística é usada em diversos campos científicos para a análise de dados binários (Agresti, 2012; Hilbe, 2009).

O modelo de regressão logística binária

Considere-se um elemento genérico duma população com características ($X = x, Y = y$). Suponha-se que Y é uma variável aleatória que apresenta distribuição Bernoulli, $Y \frown B(p)$, com função massa de probabilidade: $f(y,p) = P(Y = y) = p^y(1-p)^{1-y}, y \in 0,1$, e valor médio $E(Y) = 0 \times P(Y = 0) + 1 \times P(Y = 1) = P(Y = 1) = p$, sendo p a probabilidade de ocorrência do evento de sucesso ($Y = 1$).

Suponha-se agora que se recolhe uma amostra de dimensão n desta população. O vetor $(y_1, y_2, \cdots, y_n)$ consiste, deste modo, nas n observações independentes de um vetor aleatório constituído por variáveis aleatórias binárias independentes $(Y_1, Y_2, \cdots, Y_n)$, tal que cada $Y_i \frown B(p_i)$.

Uma vez que as observações são independentes, tem-se para função massa de probabilidade conjunta de $y_1, y_2, \cdots, y_n$ e função de verosimilhança:

$$f(y,p) = \mathcal{L}(p;y) = \prod_{i=1}^{n} \mathcal{L}_i(p_i;y_i) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}. \tag{4.12}$$

Como suposição inicial, é razoável admitir que p_i varia entre 0 e 1 ($0 < p_i < 1$), quando x_i varia na reta real, crescendo com a variável independente de forma "logística"(curva em 'S', característica da distribuição logística), como observado no gráfico exemplo da esquerda da figura 4.3.

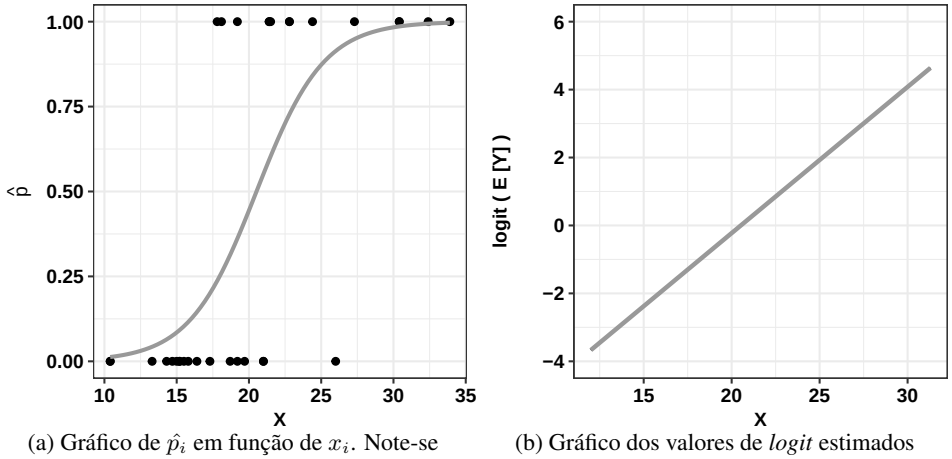
Maximizando a função de verosimilhança sem impôr qualquer tipo de restrições a p_i daria o valor degenerado $p_i = 0$ se $y_i = 0$ e $p_i = 1$ se $y_i = 1$, pelo que a probabilidade (p) é transformada em chance ($odds$, $\frac{p}{1-p}$) e é aplicada uma transformação logarítmica do $odds$, denominada de $logit$, sobre todos os possíveis valores de x de modo a que tal função de p pertença ao intervalo $]-\infty, +\infty[$:

$$logit(p_i) = \ln\left[\frac{p_i}{1-p_i}\right], \tag{4.13}$$

onde temos então que

$$\begin{aligned} & logit(p_i) = x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p \Leftrightarrow \\ & \Leftrightarrow \mu_i = p_i = \frac{exp(x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)}{1 + exp(x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)}. \end{aligned} \tag{4.14}$$

Deste modo, obtém-se um modelo de regressão logística binária, onde a ligação entre o preditor linear $x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p$ e o valor médio p_i ($E(Y_i) \equiv \mu_i = p_i$), obriga p_i a estar no intervalo $]0,1[$. Pela observação do gráfico da direita da figura 4.3 constata-se, assim, a linearidade do $logit$ com X.



(a) Gráfico de $\hat{p_i}$ em função de x_i. Note-se que $\hat{p_i}$ cresce com x_i de forma logística

(b) Gráfico dos valores de *logit* estimados para a variável X. Note-se que $logit(E(Y))$ é linear com X

Figura 4.3: Gráficos exemplo das funções logística e linear. Base de dados utilizada proveniente da biblioteca **mtcars** do *software* **R**

Estimação do modelo

A estimação dos parâmetros de um modelo de regressão logística é feita, na sua grande maioria, através do método de máxima verosimilhança, que estima os coeficientes de regressão que maximizam a probabilidade de encontrar as realizações da variável dependente observadas da amostra. Note-se que existem outros dois métodos que também permitem a estimação dos coeficientes, como o método dos mínimos quadrados ponderados e a análise de função discriminante (Hosmer Jr and Lemeshow, 2000), contudo não serão utilizados neste projeto.

Para uma manipulação algébrica de menor complexidade da função verosimilhança acima apresentada, é-lhe primeiramente aplicado o logaritmo natural, obtendo-se a função log-verosimilhança:

$$\ln \mathcal{L}(p; y) = \ell(p; y) = \sum_{i=1}^{n} \ell_i(p_i; y_i) = \sum_{i=1}^{n} \left(y_i \ln \frac{p_i}{1-p_i} + \ln(1-p_i) \right). \tag{4.15}$$

O próximo passo será encontrar os valores de $\beta_0, \beta_1, \cdots, \beta_p$ que maximizem a função de log-verosimilhança. Para tal são determinadas as derivadas parciais em ordem a $\beta_0, \beta_1, \cdots, \beta_p$ da função log-verosimilhança,

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \frac{\partial \sum_{i=1}^{n} y_i(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p) - \ln\left(1 + exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)\right)}{\partial \beta_j} \Leftrightarrow \\ \Leftrightarrow \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial p_i} \frac{\partial p_i}{\partial \beta_j} &= \frac{y_i - p_i}{p_i(1-p_i)} \frac{\partial p_i}{\partial \beta_j} = x_{ij} \frac{y_i - p_i}{p_i(1-p_i)} p_i(1-p_i) = x_{ij}(y_i - p_i), \end{aligned} \tag{4.16}$$

e posteriormente igualadas a zero, obtendo-se o seguinte sistema de equações:

$$\begin{cases} \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_0} = \sum_{i=1}^{n} x_{i0}(y_i - p_i) = 0 \\ \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_1} = \sum_{i=1}^{n} x_{i1}(y_i - p_i) = 0 \\ \vdots \\ \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_p} = \sum_{i=1}^{n} x_{ip}(y_i - p_i) = 0, \end{cases} \tag{4.17}$$

com $j = 0, 1, \cdots, p$ e $x_{i0} = 1$.

Considerando então

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}; \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}; \quad \boldsymbol{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_p \end{bmatrix},$$

este sistema denota-se matricialmente por:

$$\mathbf{X}^\top (\mathbf{Y} - \mathbf{p}) = \mathbf{0}. \tag{4.18}$$

Dado não existir uma solução analítica para este sistema de equações, torna-se necessário utilizar um método iterativo para determinar os estimadores de máxima verosimilhança (EMV). O método mais utilizado é o método iterativo dos mínimos quadrados ponderados, também conhecido por *score* de Fisher. Este método é uma adaptação do método de Newton-Raphson em que a matriz Hessiana é substituída pela quantidade de informação de Fisher (Hosmer Jr and Lemeshow, 2000).

Para inferir sobre o vetor de parâmetros $\boldsymbol{\beta}$, nomeadamente, para fazer testes de hipóteses e obter intervalos de confiança, é necessário conhecer a distribuição amostral do EMV. Para as variáveis resposta que não são consideradas normalmente distribuídas, como acontece na regressão logística binária, é necessário recorrer a resultados assintóticos baseados no Teorema Limite Central (TLC), que se verificam para grandes amostras quando os modelos em estudo satisfazem certas condições de regularidade. Em Fahrmeir e Kaufmann (1985) são estabelecidas condições que garantem a consistência e a normalidade assintótica do estimador de máxima verosimilhança, $\boldsymbol{\beta}$, dos parâmetros dos MLG (Fahrmeir and Kaufmann, 1985).

Sob as condições de regularidade e considerando uma amostra grande, podem ser deduzidas algumas propriedades assintóticas do EMV de $\boldsymbol{\beta}$:

1. $E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx E\left(\boldsymbol{I}^{-1}(\boldsymbol{\beta})\, \boldsymbol{U}\right) = 0$, isto é, $\hat{\boldsymbol{\beta}}$ é um estimador de $\boldsymbol{\beta}$ assintoticamente centrado (consistência).
2. $Cov(\hat{\boldsymbol{\beta}}) \approx E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right] = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, isto é, a variância assintótica do estimador de máxima verosimilhança corresponde ao inverso da matriz de informação de Fisher ($\boldsymbol{I}(\boldsymbol{\beta})$) calculada para o verdadeiro valor do parâmetro. É importante reconhecer que para o modelo de regressão logística se tem que

$$\boldsymbol{I}(\boldsymbol{\beta}) = E\left[-\partial^2 \ell(\boldsymbol{\beta})\right] = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} = -\partial^2 \ell(\boldsymbol{\beta}), \tag{4.19}$$

o que se traduz na equivalência dos métodos de Newton-Raphson e de *score* de Fisher, e onde $\boldsymbol{W}$ é a matriz diagonal de covariâncias ($n \times n$) obtida por

$$\boldsymbol{W} = \begin{bmatrix} p_1(1-p_1) & & & 0 \\ & p_2(1-p_2) & & \\ & & \ddots & \\ 0 & & & p_n(1-p_n) \end{bmatrix}. \tag{4.20}$$

3. A distribuição assintótica de $\hat{\boldsymbol{\beta}}$ é normal multivariada com vetor médio $\boldsymbol{\beta}$ e matriz de covariâncias $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$, ou seja

$$\hat{\boldsymbol{\beta}} \overset{n\to\infty}{\sim} N_{p+1}\left(\boldsymbol{\beta}, \boldsymbol{I}^{-1}(\boldsymbol{\beta})\right). \tag{4.21}$$

Na prática, uma vez que o vetor $\boldsymbol{\beta}$ é desconhecido, então $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ também o é, logo substitui-se $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ pela matriz de informação de Fisher calculada para a estimativa de $\hat{\boldsymbol{\beta}}$.

Seleção de Variáveis

Antes do início da construção do modelo é importante ter em conta o objetivo da análise, que consiste, neste projeto, exclusivamente em compreender a relação entre os padrões alimentares e a DMI através da estrutura dum modelo, ou seja, pretende-se construir um modelo explicativo e não um modelo preditivo. Os modelos explicativos são utilizados para explicar fenómenos ou testar explicações causais, enquanto que um modelo preditivo, por sua vez, procura produzir expectativas de comportamento futuro ou predizer futuros eventos de interesse e tendências (Shmueli et al., 2010).

As variáveis `pc1`, `pc2` e `pc3` representam as variáveis de interesse neste estudo obtidas por ACP, cujo efeito à sua exposição na variável resposta `dmi` nos interessa avaliar. Já as restantes variáveis, uma vez que consistem em potenciais fatores de risco para a doença, existindo evidências na literatura, apesar de não consistentes, sobre a respetiva associação com a doença (revisão exaustiva realizada, apresentada na secção 5.3), representam potenciais fatores de confundimento, que poderão levar à distorção da magnitude da relação entre as variáveis de interesse e a variável resposta.

A regressão logística é muitas vezes utilizada para o controlo do confundimento gerado por estas variáveis, através da sua inclusão no modelo (dos Santos Silva, 1999). Vários epidemiologistas sugerem até a inclusão de todas as variáveis que possam ser relevantes no modelo de regressão logística (Hosmer Jr and Lemeshow, 2000).

No entanto, tal como neste projeto, outros estudos epidemiológicos dispõem de um elevado número de variáveis e a inclusão de todas poderá levar à produção de estimativas numericamente instáveis e conduzir a uma maior dependência do modelo nos dados observados, um fenómeno designado de sobreajustamento (*overfitting*) (Hosmer Jr and Lemeshow, 2000). Existem diversos métodos de seleção de covariáveis com o objetivo de obter o "melhor"modelo dentro do contexto científico do problema. Entre elas, encontram-se a eliminação *backward* (seleção do melhor subconjunto do modelo completo), a seleção *forward* (seleção sequencial de variáveis independentes) ou a eliminação bidirecional

(combinação dos dois métodos anteriores) (Gomes, 2011a).

O confundimento residual consiste também noutra preocupação no processo de modelação, ocorrendo como resultado da incorreta modelação do efeito das variáveis de confundimento na variável resposta, o que leva igualmente à distorção da relação entre as variáveis de interesse e a variável resposta. Em muitos estudos, encontra-se reportada a categorização das variáveis contínuas (Chiu et al., 2014; Islam et al., 2014), abordagem esta que presume que a relação entre o preditor e a resposta é uniforme dentro dos intervalos definidos e altamente desaconselhada por vários autores, pelos problemas que dela derivam (Groenwold et al., 2013; Royston et al., 2006).

Na construção dos modelos de regressão logística binária exige-se a linearidade do $logit$ para as variáveis contínuas. No entanto, a relação entre a variável resposta e a variável independente muitas vezes é de natureza não linear, sendo a simples inclusão da variável independente insuficiente para modelar a relação (Groenwold et al., 2013; Binder et al., 2013; Benedetti and Abrahamowicz, 2004). A regressão de *splines* oferece uma forma conveniente de analisar a relação funcional, podendo ser utilizada em qualquer modelo de regressão que especifique uma função da variável resposta como função da combinação linear das variáveis independentes.

Uma função *spline* consiste numa função matemática definida por intervalos, delimitados por uma sequência de pontos, em que a cada intervalo corresponde um polinómio. Uma função *spline* de ordem $(m + 1)$ é, portanto, uma função definida seccionalmente, por polinómios de grau m, com $m - 1$ derivadas contínuas nos nós (Reis dos Santos and Reis dos Santos, 2012) (mais detalhes sobre este tópico poderão ser encontrados em (De Boor, 1978).

Existem várias técnicas de construção de modelos baseadas em *splines*, no entanto, nenhuma foi considerada como "melhor"escolha (Ruppert et al., 2003). Deste modo, neste trabalho optou-se pela técnica de *restricted cubic splines* (RCS), popularizada por (Harrell, 2014) e disponível na biblioteca **rms** do **R** (Harrell Jr, 2017). Uma regressão com *restricted cubic splines* é definida como uma função cúbica entre membros adjacentes de um conjunto de nós (*knots*) fixos $t_1 < t_2 < \cdots < t_k$ no domínio de uma variável independente X, sendo uma função linear se $x < t_1$ ou $x > t_k$, contínua e tendo primeiras e segundas derivadas contínuas. Normalmente, um número reduzido de nós (3 a 5) é suficiente para modelar a maioria dos dados. No início da presente análise optou-se por utilizar 5 nós (Binder et al., 2013), localizados nos quantis das correspondentes variáveis, segundo definido na tabela 4.1.

Comparativamente aos *splines* cúbicos, que têm pobre comportamento nas duas caudas (antes do primeiro nó e depois do último nó), a técnica *restricted cubic splines* providencia um melhor ajustamento aos dados e também reduz os graus de liberdade, uma vez que os *splines* são forçados a ser lineares nas caudas.

Tabela 4.1: Quantis de localização dos nós, segundo o número de nós definido

Nº nós	**Quantis**						
3			0.10	0.50	0.90		
4			0.05	0.35	0.65	0.95	
5		0.05	0.275	0.50	0.725	0.95	
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.50	0.6583	0.8167	0.975

Em termos de funções de X a serem incluídas como variáveis independentes, é fácil expressar uma regressão com *restricted cubic splines*, que para 5 nós implica a adição de três variáveis adicionais, aqui denominadas de $S_{5,1}$, $S_{5,2}$ e $S_{5,3}$, e definidas por:

$$\begin{array}{l} S_{5,1} = (X-t_1)_+^3 - \frac{t_5-t_1}{t_5-t_4}(X-t_4)_+^3 + \frac{t_4-t_1}{t_5-t_4}(X-t_5)_+^3 \\ S_{5,2} = (X-t_2)_+^3 - \frac{t_5-t_2}{t_5-t_4}(X-t_4)_+^3 + \frac{t_4-t_2}{t_5-t_4}(X-t_5)_+^3 \\ S_{5,3} = (X-t_3)_+^3 - \frac{t_5-t_3}{t_5-t_4}(X-t_4)_+^3 + \frac{t_4-t_3}{t_5-t_4}(X-t_5)_+^3 \end{array} , \tag{4.22}$$

onde $(X-t_z)_+$ é igual a $X-t_z$ se $X-t_z > 0$, e $(X-t_z)_+$ igual a 0 caso contrário, onde $z = 1, 2, \cdots, 5$ representa o índice dos nós.

Uma vez que nenhum método particular se encontra amplamente recomendado para a seleção de variáveis em modelos com *splines*, neste projeto procurou-se desenvolver uma estratégia que permitisse, por um lado, a seleção das variáveis de confundimento, e por outro, a seleção das formas funcionais das variáveis contínuas presentes no modelo.

A estratégia aqui aplicada resultou de uma combinação das metodologias propostas em *Multivariable model-building with continuous covariates: 2. Comparison between splines and fractional polynomials* (Binder et al., 2013) e *Using generalized additive models to reduce residual confounding* (Benedetti and Abrahamowicz, 2004), sendo constituída pelos seguintes passos:

1. **Construção do modelo inicial.** Como ponto de partida, foi construído e ajustado um modelo inicial contendo todas as covariáveis candidatas, com uma componente RCS com 5 nós para cada covariável contínua.

2. **Testar não linearidade das variáveis de interesse.** Entre as variáveis de interesse, é identificada a variável com menor importância no modelo, ou seja, a que apresente o maior valor-p no teste de razão de verosimilhanças (descrito mais à frente nesta secção) entre o modelo corrente e o modelo sem essa variável.

   Inicialmente foram construídos vários modelos com o número de nós da componente RCS dessa variável de interesse a variar entre 2 a 9, de modo a determinar o número ótimo de nós a utilizar. Os valores do critério de informação de Akaike (*Akaike information criterion*, AIC), critério esse que se encontra igualmente descrito mais à frente nesta secção, foram utilizados para comparar estes modelos (não encaixados), onde o melhor modelo (menor AIC) foi selecionado para o subsequente teste de não linearidade (Akaike, 1974).

   Foram adicionalmente construídos modelos com a variável contínua na forma não linear polinomial (2º e 3º graus).

   Finalmente, foram realizados testes de razão de verosimilhanças entre o modelo com a variável de interesse na sua forma linear e os modelos com a variável de interesse nas diferentes formas não lineares, testes estes designados genericamente por testes de não linearidade. Caso, para qualquer dos testes realizados, o valor-p fosse superior ao nível de significância aqui definido ($\alpha = 0.05$), não haveria evidência para rejeitar a hipótese da linearidade da variável, logo o modelo seria reajustado com a variável na sua forma linear. Pelo contrário, se um ou mais testes apresentassem valor-p igual ou inferior a 0.05, os valores AIC desses modelos não lineares seriam comparados e o modelo reajustado com a variável na forma funcional com menor AIC.

Este processo foi repetido para as restantes variáveis de interesse. Salienta-se que o modelo corrente é agora o modelo reajustado. O modelo final obtido neste passo será em diante denominado de modelo completo.

3. **Seleção de variáveis de confundimento.** Este procedimento teve início com a variável de confundimento que apresentasse menor significância no modelo corrente, ou seja, com maior valor-p no teste de razão de verosimilhanças entre esse modelo e o modelo sem a variável em questão. O valor-p teria também de ser igual ou superior a 0.20, para a variável poder ser excluída, valor este sugerido por (Greenland, 2008), que por ser relativamente elevado mitiga, em parte, o problema da exclusão de variáveis de confundimento potencialmente importantes (Maldonado and Greenland, 1993).

   Os coeficientes das variáveis de interesse no modelo completo foram então comparados com os coeficientes do novo modelo ajustado sem a variável de confundimento:

$$\frac{\beta_{jc} - \beta_{js}}{\beta_{jc}} \times 100, \tag{4.23}$$

   onde β_{jc} representa o coeficiente da variável de interesse j ($j = 1, 2, 3$) no modelo completo e β_{js} os coeficientes para essa variável no modelo sem a variável de confundimento a ser testada.

   Caso alguma das variáveis de interesse verificasse uma alteração dos seus coeficientes superior a 10%, a variável de confundimento seria retida no modelo, dada a sua importância no ajustamento necessário do efeito da variável de interesse (Hosmer Jr and Lemeshow, 2000). Caso contrário, um novo modelo seria reajustado sem essa variável.

   Este procedimento foi repetido para as todas as restantes variáveis de confundimento, ou até alguma apresentar um valor-p no teste de razão de verosimilhanças inicial igual ou superior a 0.20.

4. **Testar não linearidade das variáveis de confundimento retidas.** O procedimento descrito em **2.** foi repetido para todas as variáveis de confundimento que permaneceram no modelo obtido em **3.**. Para as variáveis que demonstraram ter uma forma funcional não linear, foram adicionalmente comparados os coeficientes das variáveis de interesse do modelo com a variável de confundimento na sua forma linear e do modelo com essa variável na forma não linear selecionada. Caso a alteração de agum desses coeficientes fosse superior a 10%, optar-se-ia pela forma funcional não linear, sendo o modelo reajustado.

5. **Introdução de interações.** O último passo consistiu na determinação da existência de interações entre as covariáveis do modelo. Uma interação entre duas variáveis implica que o efeito de uma não seja constante ao longo dos vários níveis da outra variável. Além disso, todas as interações incluídas deverão ser clinicamente relevantes, devendo ser feita inicialmente uma lista de interações plausíveis a nível clínico. As interações clinicamente relevantes são então adicionadas, uma de cada vez, ao modelo e é avaliada a sua significância com um teste de razão de verosimilhanças. Deverão ser mantidas apenas as que apresentarem um valor-p igual ou inferior a 0.05 e as que não gerem grandes alterações nos valores dos restantes coeficientes do modelo.

Testes estatísticos utilizados na seleção de variáveis

No processo de seleção de variáveis, a decisão relativa ao teste estatístico a utilizar para comparar dois modelos vai depender dos modelos estarem ou não encaixados. O teste de razão de verosimilhanças

(teste RV) é o indicado para comparar dois modelos encaixados estimados sobre o mesmo conjunto de dados. Testa-se a nulidade de um subvetor r de componentes de $\boldsymbol{\beta}$,

$$H_0 : \boldsymbol{\beta}_r = 0 \quad vs \quad H_1 : \boldsymbol{\beta}_r \neq 0,$$

onde a estatística de teste é dada por

$$-2\ln(\mathcal{L}_s/\mathcal{L}_c) = -2(\ln \mathcal{L}_s - \ln \mathcal{L}_c) \underset{H_0}{\sim} \chi^2_{k_c - k_s}, \tag{4.24}$$

em que $\mathcal{L}_c$ corresponde à verosimilhança do modelo mais geral, com k_c parâmetros e $\mathcal{L}_s$ corresponde à verosimilhança do modelo encaixado (mais simples) com k_s parâmetros. Sob a hipótese nula de que o modelo restrito é mais adequado (ou seja, de que os $r = k_c - k_s$ parâmetros adicionais são nulos) a estatística de teste tem distribuição assintótica de um qui-quadrado com $k_c - k_s$ graus de liberdade.

Outro teste utilizado nas análises estatisticamente equivalente ao teste RV, é o teste de Wald (Hosmer Jr and Lemeshow, 2000). Para avaliar a significância de apenas um determinado coeficiente j ($j = 1, 2, \cdots, p$) do modelo, onde se testa

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0,$$

sendo a estatística de teste

$$W = \frac{\hat{\beta}_j}{\sqrt{se(\hat{\beta}_j)}}, \tag{4.25}$$

que sob H_0 tem distribuição assintótica gaussiana, com base nas propriedades dos estimadores para $\boldsymbol{\beta}$. Note-se que se o efeito de β_j não for significativo, a variável X_j associada não representa, portanto, um preditor importante para a variável resposta.

No caso multiparamétrico, admitindo que o modelo contém p parâmetros além de β_0, pretende-se testar a hipótese

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \underline{0} \quad vs \quad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \underline{0},$$

em que $\boldsymbol{C}$ é uma matriz de $r \times p$ de característica completa r, por exemplo

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 1 & \dots & 0 \end{bmatrix},$$

que nos permite testar r hipóteses sob a forma de um sistema de equações. No exemplo da matriz $\mathbf{C}$ acima fornecida, pretende-se, portanto, testar as hipóteses $\beta_2 = 0$ e $\beta_3 + \beta_4 = 0$.

A estatística de teste é então

$$W = \boldsymbol{C}\hat{\boldsymbol{\beta}}^\top [\boldsymbol{C}\ var(\hat{\boldsymbol{\beta}})\boldsymbol{C}^\top]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}) \underset{H_0}{\sim} \chi^2_r, \tag{4.26}$$

que sob H_0 apresenta distribuição assintótica de um qui-quadrado com r graus de liberdade.

Quando estamos perante dois modelos em que nenhum pode ser representado como um caso especial do outro, podemos classificá-los como modelos não encaixados. Quando tal se verifica, é utilizado o AIC para a comparação desses modelos. O valor de AIC representa uma medida relativa da perda de informação nesse modelo particular, sendo que o melhor modelo deverá ter menor AIC. Este valor é calculado por:

$$AIC = -2\ln\mathcal{L} + 2k, \tag{4.27}$$

onde k representa o número de parâmetros e $\mathcal{L}$ a verosimilhança do modelo. Note-se que este critério penaliza o número de parâmetros do modelo. Apresenta também a desvantagem de não permitir testar hipóteses, não fornecendo informação sobre a qualidade do modelo.

Avaliação do modelo ajustado

Após a estimação dos parâmetros do modelo, existem vários passos envolvidos na avaliação do ajustamento, adequabilidade e utilidade do modelo. A qualidade geral do ajustamento do modelo (*goodness-of-fit*) envolve determinar se a variação dos resíduos do modelo ajustado é pequena, não dispõe de tendências sistemáticas e segue a variabilidade postulada pelo modelo (Hosmer et al., 1997). A violação de uma ou mais destas três características, pode resultar na falta de ajustamento do modelo.

A qualidade geral do ajustamento do modelo pode ser verificada através do valor de AIC do modelo e da aplicação de testes mais gerais que investigam quão próximos são os valores preditos pelo modelo proposto dos valores observados, como por exemplo o teste de Hosmer-Lemeshow, frequentemente utilizado em regressão logística, o teste de qui-quadrado de Pearson e o teste da *Deviance* (descrição detalhada dos métodos em (Hosmer and Lemesbow, 1980)). No entanto, estes métodos clássicos apresentam várias desvantagens: os resultados do teste de Hosmer-Lemeshow dependem substancialmente do número de grupos definidos, e não existe uma regra que guie a decisão desse número; as distribuições das estatísticas de teste de Pearson e da *Deviance* poderão ser consideravelmente diferentes de uma distribuição qui-quadrado quando utilizadas variáveis quantitativas no ajustamento do modelo, pois tal resulta muitas vezes em apenas um caso por perfil, que se define como um grupo de casos que observam exatamente os mesmos valores para as variáveis preditoras (Allison, 2014).

Por estes motivos, neste trabalho foi aplicado um teste que pode ser calculado com apenas um caso por perfil e sem agrupamento de observações, denominado teste de soma não ponderada de quadrados (Allison, 2014). Este foi proposto em 1989 por Copas para testar proporções (Copas, 1989). Aqui, foi simplicada a estatística de teste original de Copas e providenciada a sua distribuição assintótica sob a situação estritamente binária, como sugerido em (Wu, 2010).

A estatística de teste é

$$S = \sum_{i=1}^{n}(y_i - \hat{p}_i)^2, \tag{4.28}$$

que sob a hipótese nula de que o modelo de regressão ajustado é correto em todos os aspetos, os momentos assintóticos de $\hat{S}$ são $E[\hat{S} - tr(\boldsymbol{W})] \cong 0$ e $Var(\hat{S} - tr(\boldsymbol{W})) \cong \boldsymbol{d}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{W}\boldsymbol{d}$. Aqui, $\boldsymbol{d}$ é o vetor com elemento geral $d_i = 1 - 2p_i$ e $\boldsymbol{W} = diag[p_i(1 - p_i)]$ é a matriz de covariâncias $n \times n$. A matriz $\boldsymbol{H}$ (*hat matrix*) de dimensão $n \times n$, no contexto de regressão logística, é conhecida como a matriz *Pregibon leverage*, sendo obtida por

$$\boldsymbol{H} = \hat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{W}}^{1/2}, \tag{4.29}$$

onde $\boldsymbol{X}$ é a matriz de *design* $n \times (k+1)$, contendo as medidas observadas das k variáveis para os n indivíduos do estudo e uma coluna (a primeira) apenas com o valor 1 para todos os indivíduos.

Deste modo, após a substituição de p_i pelas respetivas estimativas, a estatística estandardizada

$$\frac{[\hat{S} - tr(\hat{\boldsymbol{W}})]}{\sqrt{\hat{Var}(\hat{S} - tr(\hat{\boldsymbol{W}}))}} \tag{4.30}$$

segue uma distribuição normal padrão.

Diagnóstico do modelo

As estatísticas sumárias da qualidade do modelo acima referidas providenciam um único valor que sumariza a concordância entre os valores observados e preditos. Tal traduz-se tanto numa vantagem como numa desvantagem, pois um único valor sumariza informação considerável. Antes de se concluir que o modelo se encontra bem ajustado é, portanto, crucial analisar outras medidas para perceber se o ajustamento é suportado ao longo de todo o conjunto de padrões de covariáveis, em que um padrão de covariáveis consiste num conjunto único de valores para as covariáveis no modelo.

Um conjunto de medidas especializadas é aplicado neste contexto, constituindo, de modo geral, na base do diagnóstico do modelo.

Multicolinearidade

A multicolinearidade (ou colinearidade) ocorre quando duas ou mais variáveis independentes no modelo são aproximadamente determinadas por uma combinação linear de outras variáveis independentes no modelo, ou seja, quando existe uma grande correlação entre essas. Tal fenómeno pode resultar numa grande variabilidade nas estimativas dos coeficientes de regressão, tornando o modelo menos significativo e menos robusto. A existência de multicolinearidade pode ser avaliada através do cálculo do *Variance Inflation Factor* (VIF).

O VIF permite medir o quanto a variância da estimativa dos coeficientes se encontra inflacionada comparativamente a quando as variáveis não são linearmente dependentes. No entanto, esta medida é apenas aplicável a modelos com termos de apenas um coeficiente, como preditores quantitativos com efeitos lineares. Termos de múltiplos coeficientes, tais como conjuntos de regressores *dummy* representando um preditor categórico, requerem uma abordagem mais geral, dado que as correlações entre regressores, num conjunto relacionado, são afetadas por mudanças no modelo, como a mudança na categoria de referência para o conjunto de regressores *dummy*.

Em 1992, John Fox e Georges Monette propuseram o *Generalized Variance Inflation Factor* (GVIF) tendo em conta esses casos específicos (Fox and Monette, 1992). Enquanto que os VIF's representam o impacto da colinearidade no quadrado da amplitude do intervalo de confiança para um coeficiente, os GVIF's medem o impacto no quadrado da tamanho da região de confiança para p coeficientes[a]:

$$GVIF_1 = \frac{det\boldsymbol{R_{11}}\, det\boldsymbol{R_{22}}}{det\boldsymbol{R}}, \tag{4.31}$$

onde $\boldsymbol{R_{11}}$ é a matriz de correlação entre os regressores em $\boldsymbol{X_1}$;$\boldsymbol{R_{22}}$ é a matriz de correlação entre os regressores em $\boldsymbol{X_2}$; $\boldsymbol{R}$ é a matriz de correlação para $(\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_p})$; e *det* a função determinante.

[a] A uma variável preditora categórica com $p+1$ categorias correspondem p coeficientes no modelo (p graus de liberdade).

Esta medida pode ser obtida com recurso à função **vif()** da biblioteca **car** do R.

Fox e Monette sugeriram inclusivamente a elevação do GVIF a $\frac{1}{2p}$, de modo a tornar o índice comparável entre os diferentes valores de p. Ou seja, os GVIF's consistem, de maneira geral, nos VIF's corrigidos pelos graus de liberdade (p) da variável preditora, sendo que no caso específico de a variável preditora ter apenas um grau de liberdade (variáveis quantitativas ou dicotómicas), o GVIF equivale à raíz quadrada do VIF. Vários valores de VIF têm sido utilizados como limites críticos para indicar multicolinearidade excessiva, como os valores 4, 5 ou 10 (O'brien, 2007). Como tal, neste projeto optou-se pelo valor crítico do GVIF para uma variável preditora de $10^{\frac{1}{2p}}$ com fim a avaliar a respetiva colinearidade (O'brien, 2007).

Análise dos resíduos

A análise dos resíduos é a designação usada para um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos. Os resíduos (e_i) representam a discrepância entre os valor observados (Y_i) para a variável resposta e o valores ajustados ($\hat{Y}_i$), isto é

$$e_i = y_i - \hat{y}_i, \quad i = 1, \cdots, n. \tag{4.32}$$

Nos modelos de regressão lineares, este tipo de análise tem um papel particularmente importante no que respeita à avaliação da qualidade do ajustamento do modelo e à verificação das condições de Gauss-Markov e de normalidade. No entanto, a análise de resíduos para a regressão logística é mais complexa, uma vez que mesmo na presença de um modelo correto, os resíduos poderão estar correlacionados e, dada a natureza binária da variável resposta, não terem distribuição normal. Deste modo, vários autores não recomendam este tipo de análise (Kleinbaum et al., 2013; Ottenbacher et al., 2001).

Neste projeto, a análise dos resíduos focou-se apenas na deteção de potenciais *outliers* e observações influentes, uma tarefa crucial no processo de modelação, dado que essas observações podem levar a uma distorção da validade e adequabilidade do modelo. Para os modelos de regressão logística binária, vários autores recomendam que a abordagem da deteção dos *outliers* seja baseada nos resíduos estandardizados e nos resíduos $deviance$, uma vez que perante uma amostra de larga escala, esses apresentam uma distribuição mais próxima da normal padrão quando o modelo é correto. Deste modo, serão apenas descritos nesta secção e utilizados na análise de dados estes dois tipos de resíduos, apesar de outros se encontrarem citados na literatura (Hosmer Jr and Lemeshow, 2000).

Primeiramente, comece-se por definir o i-ésimo resíduo ordinário, que poderá apenas assumir um de dois valores

$$e_i = \begin{cases} 1 - \hat{p}_i, & \text{se } Y_i = 1 \\ -\hat{p}_i, & \text{se } Y_i = 0 \end{cases} \tag{4.33}$$

Os resíduos de Pearson (pr_i) são obtidos pela divisão dos resíduos ordinários pelo erro padrão estimado de Y_i e definidos por

$$pr_i = \frac{e_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \frac{(Y_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}. \tag{4.34}$$

Estes resíduos estão diretamente relacionados à estatística de teste de χ^2 de Pearson, uma vez que $X^2 = \sum_{i=1}^{n} e_i^2$, daí a sua denominação. Contudo, os resíduos assim obtidos poderão não ter variância unitária,

consequentemente, os resíduos ordinários são ainda estandardizados pelos seus respetivos erros padrão estimados, aproximados por $\sqrt{\hat{p}_i(1-\hat{p}_i)(1-h_{ii})}$, sendo obtidos os denominados resíduos de Pearson estandardizados ou estudantizados (prs_i). Estes são definidos por

$$prs_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-h_{ii})}} = \frac{pr_i}{\sqrt{1-h_{ii}}}, \tag{4.35}$$

onde h_{ii} representa o i-ésimo elemento da diagonal da matriz $n \times n$ **H** (ver secção 4.4). Como mencionado, esta é também conhecida como a matriz *Pregibon leverage*, uma vez que mede os *leverage* ou valores de alavancagem de uma observação.

Os valores de alavancagem são uma medida da importância de uma observação no ajustamento do modelo, variando de 0 a 1, onde o valor de alavancagem 1 significa que o modelo está a ser forçado a ajustar essa observação de forma exata, sendo essa, portanto, de grande influência. O limite de $2(k+1)/n$ tem sido proposto (Bagheri et al., 2010; Belsley et al., 2013), em que observações com h_{ii} superior são declaradas como influentes. Note-se que os valores dos resíduos de Pearson estandardizados crescem com os valores de alavancagem. Deste modo, este tipo de resíduos, ao contrário dos resíduos de Pearson, são ferramentas importantes na identificação de observações influentes. Além disso, para uma amostra grande ($n \geq$30) seguem aproximadamente uma distribuição normal padrão.

Outro tipo de resíduo bastante útil na identificação de *outliers* ou observações mal ajustadas no modelo são os resíduos *deviance* (dr_i). Estes medem a discordância entre qualquer uma das componentes da log-verosimilhança do modelo ajustado e a componente correspondente da log-verosimilhança que iria ser obtida se cada ponto fosse ajustado de forma exata. O objetivo da regressão logística é, deste modo, minimizar a soma dos resíduos *deviance*, definidos por

$$dr_i = sinal(Y_i - \hat{p}_i)\left\{-2[Y_i - \ln(\hat{p}_i) + (1-Y_i)\ln(1-\hat{p}_i)]\right\}^{\frac{1}{2}}. \tag{4.36}$$

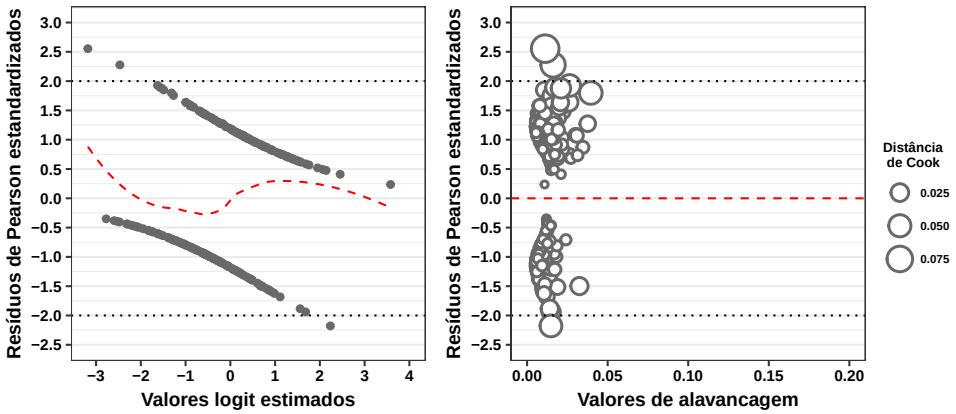
Estes resíduos apresentam uma distribuição mais aproximada da normal que os resíduos de Pearson, uma propriedade desejável para os resíduos, daí que sejam frequentemente preferidos.

Os resíduos de Pearson estudantizados, os resíduos *deviance* e os valores de alavancagem estão, portanto, nas bases do diagnóstico da regressão logística binária.

Em termos práticos, uma boa maneira de avaliar o impacto dos resíduos consiste na sua representação gráfica contra as probabilidades logísticas estimadas. Os gráficos resultantes produzem sempre padrões similares: duas linhas de tendência com declive aproximadamente -1, dado que os resíduos adoptam apenas um de dois valores num ponto X_i, $1-\hat{p}_i$ ou $0-\hat{p}_i$ (ver gráfico da esquerda da figura 4.4). A observação deste tipo de gráficos permite a identificação de resíduos estandardizados que, se tiverem valor absoluto superior a 2 (regra dos 2-σ), são considerados potenciais *outliers*, merecendo uma inspeção detalhada. No gráfico apresentado como exemplo, podemos verificar que existem 3 observações cujos resíduos de Pearson estandardizados excedem o valor absoluto 2 e também outras observações próximas desse valor, pelo que são consideradas potenciais *outliers*.

Estes gráficos são também suplementados com uma curva *lowess smooth* dos resíduos contra as probabilidades logísticas estimadas. A ferramenta estatística *LOWESS* (*Locally Weighted Scatterplot Smoothing*) é bastante utilizada em análise de regressão, criando uma linha suave ao longo de um gráfico de dispersão para ajudar a visualizar e analisar a relação entre as variáveis e prever tendências (Sarkar et al., 2011). No presente caso, deverá produzir uma linha aproximadamente horizontal com interseção

em zero, dado que se o modelo estiver correto, então $E(Y_i - p_i)$, seguindo assintoticamente $E(Y_i - \hat{p}_i) = E(e_i) = 0$. Caso tal não se verifique, o modelo poderá ser inadequado e os potenciais *outliers* poderão estar a ter um papel relevante no ajustamento do modelo. No exemplo apresentado, a curva *lowess* pode ser considerada aproximadamente horizontal, podendo-se concluir que não existem inadequações significativas no modelo e *outliers* influentes no espaço covariado.



(a) Gráfico dos resíduos de Pearson estandardizados em função dos valores *logit* estimados. A linha tracejada vermelha representa a curva *lowess smooth* dos resíduos contra as probabilidades logísticas estimadas.

(b) Gráfico de influência proporcional dos resíduos de Pearson estandardizados em função dos valores de alavancagem. O tamanho dos círculos é proporcional à distância de Cook apresentada para as respetivas observações.

Figura 4.4: Gráficos exemplo utilizados no diagnóstico de um modelo de regressão logística binária. Base de dados utilizada disponível no Website **Freakonometrics**, acessado em 15 Fevereiro de 2017, URL **http://freakonometrics.free.fr/probit.R**

Uma vez que estes gráficos originam pouca informação sobre os *outliers* influentes, outras medidas de influência são usadas de modo complementar na identificação de observações influentes relativamente à estimação dos coeficientes de regressão logística. Entre elas, uma frequentemente usada consiste numa aproximação da distância de Cook para os MLG. Esta é obtida como a diferença estandardizada entre $\hat{\beta}$ e $\hat{\beta}_{(-i)}$, que representam as estimativas de máxima verosimilhança baseadas no conjunto de dados completo e excluindo a i-ésima observação, respetivamente, com posterior estandardização pela matriz de covariâncias estimada de $\hat{\beta}$,

$$\Delta\hat{\beta}_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^\top \, (\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X}) \, (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}) = \frac{spr_i^2 \, h_{ii}}{1 - h_{ii}}. \tag{4.37}$$

Existe um grande número de gráficos de diagnóstico sugeridos para detetar *outliers* e observações influentes. Neste projeto irão somente ser utilizados gráficos de bolhas (*bubble plots*) ou de influência proporcional, onde o valor dos resíduos de Pearson estandardizados será representado contra os valores de alavancagem, sendo que o tamanho dos símbolos circulares será proporcional à magnitude de $\Delta\hat{\beta}_i$, ou distância de Cook (ver gráfico da direita da figura 4.4).

A análise dos gráficos descritos será complementada com uma tabela com os valores dos resíduos, de alavancagem e distâncias de Cook. As observações com resíduos de valor absoluto superior a 2 serão consideradas potenciais *outliers*. Estas observações e/ou aquelas com maiores valores de alavancagem e distância de Cook serão removidas individualmente ou em grupo, e será analisada a alteração nos coeficientes do modelo. Se for elevada, as observações serão removidas, individualmente ou em grupo, e um novo modelo será ajustado.

Interpretação do modelo

Nos modelos de regressão logística, a magnitude do efeito de uma variável independente X_j ($j = 1, 2, \cdots, p$) na variável resposta pode ser descrita por $exp(\beta_j)$, dado demonstrar-se que

$$exp(\beta_j) = OR_j, \tag{4.38}$$

onde OR_j representa a razão de chances ($odds - ratio$, OR) para X_j ajustado para as outras variáveis explicativas. O OR mede a força da associação entre a variável dependente e qualquer variável explicativa, depois de descontado o efeito das outras variáveis do modelo e verificada a sua significância.

No entanto, a expressão 4.38 só é válida para modelos sem interação. Caso o modelo contenha alguma interação, o OR de uma variável depende dos valores de outras variáveis explicativas, ou seja, é impossível descrever o efeito de uma variável através de um único valor de OR.

No caso de a variável explicativa X_j ser contínua, num modelo sem interações o OR_j descreve o fator pelo qual o $odds$ de um evento muda para cada unidade de aumento de X_j. Já uma alteração de d unidades em x_j ($d = x_j'' - x_j'$) é dada, em termos de *odds ratio*, por $exp(d\ \beta_j)$.

Um intervalo de confiança para $exp(d\ \beta_j)$ pode ser obtido a partir do correspondente intervalo para $d\ \beta_j$ e é dado por

$$\left[exp\left\{d\ \hat{\beta}_j - \chi_Z(1-\alpha/2)\ d\ \hat{\sigma}(\hat{\beta}_j)\right\},\ exp\left\{d\ \hat{\beta}_j + \chi_Z(1-\alpha/2)\ d\ \hat{\sigma}(\hat{\beta}_j)\right\}\right], \tag{4.39}$$

onde $\chi_Z(1-\alpha/2)$ é o quantil $(1-\alpha/2)$ da normal padrão. Neste caso, o OR_j que relaciona o nível de exposição x_j'' comparativamente ao nível x_j' expressa a razão da vantagem em favor do valor positivo de Y quando $X_j = x_j''$ pela vantagem quando $X_j = x_j'$, mantendo as outras variáveis constantes.

Já no caso de a variável X_j ser dicotómica, o OR_j representa a comparação entre o $odds$ de um evento positivo ocorrer aquando da exposição a X_j ($X_j = 1$) e o $odds$ na ausência de exposição ($X_j = 0$), com as outras variáveis explicativas constantes (Scotia, 2010), sendo o respetivo intervalo de confiança obtido por

$$\left[exp\left\{\hat{\beta}_j - \chi_Z(1-\alpha/2)\ \hat{\sigma}(\hat{\beta}_j)\right\},\ exp\left\{\hat{\beta}_j + \chi_Z(1-\alpha/2)\ \hat{\sigma}(\hat{\beta}_j)\right\}\right]. \tag{4.40}$$

A interpretação e obtenção de intervalos de confiança para o OR para variáveis com k categorias ($C_1, C_2, \cdots, C_k$) é feita do mesmo modo. No entanto, define-se primeiramente uma categoria de referência (C_1, por exemplo), constroem-se variáveis $dummy$ para cada classe ou categoria, e calcula-se

$$OR(C_i|C_1) = exp(\beta_{i-1}), \tag{4.41}$$

com $i = 2, 3, \cdots, k$, comparando-se assim o $odds$ da classe i com o $odds$ da classe 1.

5 Resultados

Neste capítulo serão apresentados os resultados obtidos através das várias metodologias descritas no capítulo anterior. Primeiramente, na secção 5.1 serão comparados os indivíduos excluídos e incluídos quanto à prevalência de DMI e às características demográficas. Serão também comparados os indivíduos com e sem DMI incluídos no estudo quanto às características demográficas.

Proceder-se-à à identificação dos padrões alimentares, na secção 5.2, onde na subsecção 5.2.1 se encontram os resultados obtidos da aplicação da análise de componentes principais, e descritos os padrões alimentares retidos para a população em estudo. Na subsecção 5.2.2 serão apresentados os resultados do procedimento realizado para a determinação do número ótimo de *clusters* e do método de classificação não hierárquica a utilizar e da validação da solução final obtida. A subsecção 5.2.3 ir-se-á focar na comparação dos padrões alimentares obtidos com análise em componentes principais e análise classificatória, onde será feita uma caracterização de cada *cluster* obtido, acompanhada de uma breve discussão sobre a sua semelhança com os padrões derivados por ACP. Por fim, nas subsecções 5.2.4 e 5.2.5, serão feitas comparações das características demográficas e da ingestão média diária de vários nutrientes por parte dos indivíduos do estudo com diferentes adesões a cada padrão e entre os vários padrões alimentares, respetivamente.

A secção 5.3 ir-se-á concentrar na avaliação da associação entre os padrões alimentares e o risco de DMI. Inicialmente, será revista alguma literatura, de modo a selecionar as potenciais variáveis de confundimento e, assim, justificar a sua introdução nos modelos de regressão logística binária construídos. De seguida, serão apresentados os resultados das várias etapas da construção do modelo final e, adicionalmente, da respetiva avaliação de ajustamento e diagnóstico.

5.1 Análise exploratória

Dos 999 participantes do estudo, 985 (98.6 %) tinham fotografias retinianas classificáveis para DMI (ver diagrama de fluxo CONSORT na figura 7.1, Apêndice C). Após a exclusão dos indivíduos nos extremos 1% de consumo energético ($n = 18$) em cada sexo (ver secção 2.4, figura 2.1), os restantes 967 (96.8%) participantes foram incluídos na análise. Dos participantes incluídos, 428 apresentaram DMI (44.26%) e entre os participantes excluídos que tiveram fotografias retinianas classificáveis para DMI ($n = 18$), 6 apresentaram DMI (33.33%). Utilizando o teste de homogeneidade do qui-quadrado, verifica-se que não existe diferença significativa aos níveis de significância usuais na distribuição de DMI entre a população dos indivíduos incluídos e a população dos indivíduos excluídos (valor-$p = 0.355$).

Tabela 5.1: Características demográficas dos indivíduos incluídos e excluídos

	Incluídos $n = 967$	**Excluídos** $n = 18$	**valor-*p*** [a]
Demografia			
Idade, média (DP)	69.53 (7.77)	68.36 (8.27)	0.427
Sexo masculino, n (%)	431 (44.57)	8 (44.44)	1
Escolaridade			
1 - 6 anos, n (%)	174 (17.99)	3 (16.67)	
7 - 9 anos, n (%)	639 (66.08)	11 (61.11)	0.681
$\geq$ 10 anos, n (%)	154 (15.93)	4 (22.22)	
Biometria			
IMC, média (DP)	28.54 (4.36)	28.27 (4.83)	0.672
Perímetro abdominal, média (DP)	98.82 (13.70)	98.50 (11.46)	0.915
Estilo de vida			
Fumador ou ex-fumador, n (%)	253 (26.16)	7 (38.89)	0.557
Pacotes de tabaco por ano, média (DP)	8.26 (21.94)	12.76 (22.14)	0.117
Actividade física regular, n (%)	281 (29.06)	5 (27.78)	1
Consumo energético total, média (DP)	1934.99 (499.94)	2474.19 (2269.48)	0.994
Consumo de álcool, média (DP)	10.61 (14)	20.14 (33.81)	0.992
Comorbilidades			
Diabetes, n (%)	255 (26.37)	6 (33.33)	0.590
Hipertensão, n (%)	663 (68.56)	11 (61.11)	0.676
Dislipidemia, n (%)	536 (55.43)	12 (66.67)	0.477

[a] Os valores-p para as variáveis categóricas à excepção de `esc_cat`, `jafumador` e `diabetes`, foram obtidos pelo teste de homogeneidade de qui-quadrado. O teste exato de Fisher foi aplicado nas restantes variáveis, que não reuniram critérios requeridos para aplicação correta do teste qui-quadrado. Para as variáveis contínuas foi aplicado o teste não paramétrico de Mann-Whitney-Wilcoxon.

Tabela 5.2: Características demográficas dos indivíduos incluídos no estudo de acordo com a presença ou ausência de DMI

	Sem DMI $n = 539$	**Com DMI** $n = 428$	**valor-*p*** [a]
Demografia			
Idade, média (DP)	69.32 (7.65)	69.79 (7.91)	0.357
Sexo masculino, n (%)	244 (45.27)	187 (43.69)	0.671
Escolaridade			
1 - 6 anos, n (%)	91 (16.88)	83 (19.39)	
7 - 9 anos, n (%)	356 (66.05)	283 (66.12)	0.400
$\geq$ 10 anos, n (%)	92 (17.07)	62 (14.49)	
Biometria			
IMC, média (DP)	28.66 (4.33)	28.40 (4.41)	0.376
Perímetro abdominal, média (DP)	99.33 (13.03)	98.18 (14.48)	0.203
Estilo de vida			
Fumador ou ex-fumador, n (%)	144 (26.72)	109 (25.47)	0.715
Pacotes de tabaco por ano, média (DP)	8.23 (22.24)	8.30 (21.59)	0.959
Actividade física regular, n (%)	166 (30.80)	115 (26.87)	0.206
Consumo energético total, média (DP)	1958.94 (503.06)	1904.83 (494.91)	0.094
Consumo de álcool, média (DP)	10.50 (14.26)	10.76 (13.68)	0.772
Comorbilidades			
Diabetes, n (%)	150 (27.83)	105 (24.53)	0.279
Hipertensão, n (%)	361 (66.98)	302 (70.56)	0.262
Dislipidemia, n (%)	302 (56.03)	234 (54.67)	0.722

[a] Os valores-p para as variáveis categóricas foram obtidos pelo teste de homogeneidade de qui-quadrado. Para as variáveis contínuas foi aplicado o teste t-Student para amostras independentes. Para a variável `perimetroabd`, dado existir evidência estatística de não homogeneidade de variâncias (teste F com valor-$p \leq 0.05$), foi utilizada a aproximação de Welch-Satterthwaite para testar a igualdade dos valores médios das duas populações.

Na tabela 5.1 encontram-se descritas as características sociais, económicas, demográficas e respeitantes ao estilo de vida e historial médico avaliadas para os indivíduos excluídos e incluídos no estudo. Pela sua análise, verifica-se que os participantes incluídos não diferiram significativamente dos participantes excluídos quanto à distribuição dessas características.

As características sociais, económicas, demográficas e respeitantes ao estilo de vida e historial médico de acordo com a presença ou ausência de DMI encontram-se descritas na tabela 5.2. Como esperado, devido ao emparelhamento inicial no processo de seleção da amostra, as distribuições do sexo e idade entre os grupos com e sem DMI não diferiram significativamente. Relativamente às restantes características, estas também apresentaram uma distribuição semelhante nos dois grupos.

5.2 Identificação de padrões alimentares

5.2.1 Análise em componentes principais

Como referido anteriormente, um dos objetivos consistiu na identificação de padrões alimentares na população em estudo. Para tal, procedeu-se inicialmente ao agrupamento de itens alimentares do questionário semiquantitativo alimentar para minimizar variações intra-pessoais no consumo individual de alimentos, como descrito na secção 3.2. Foi então conduzida uma análise em componentes principais para derivar padrões alimentares com base nas variáveis referentes aos 26 grupos alimentares (descrição das variáveis na tabela 7.5, Apêndice B).

Com base no critério do valor próprio superior a 1 (critério de Kaiser), foram consideradas inicialmente 8 componentes principais, com a variância total explicada em 46.90%. De seguida, pela análise do *scree plot* identificaram-se as componentes com maior destaque relativamente ao valor próprio, de modo a detetar os principais padrões alimentares (figura 5.1). Foram assim identificados três padrões principais (valores próprios de 2.17, 1.87 e 1.63, respetivamente).

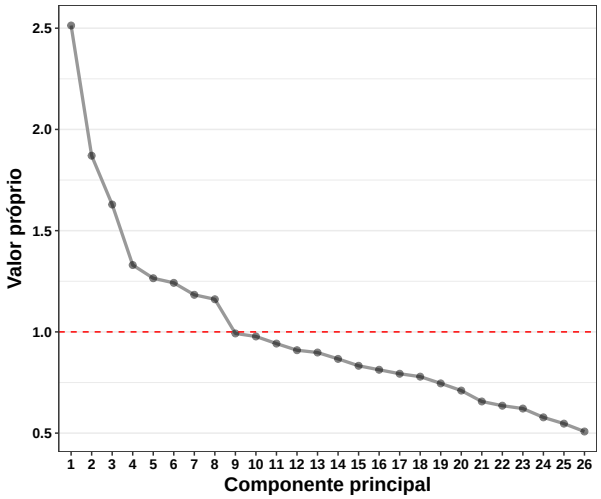


Figura 5.1: *Scree Plot* (número de componentes principais *versus* valores próprios)

De seguida procedeu-se à rotação ortogonal *varimax* das componentes extraídas Ocké (2013) e foram comparadas essas componentes sem e com rotação (ver tabela 7.8, Apêndice C). Pela análise

da tabela, verificou-se um aumento da interpretabilidade das componentes principais com a rotação *varimax*, pelo que se decidiu prosseguir a análise com as componentes rotadas.

Para a definição dos três padrões extraídos (componentes principais), foram tidos em conta apenas os grupos alimentares com *loadings* de valor absoluto superior a 0.2, que se encontram na tabela 5.3. O nome ou designação atribuída a cada padrão alimentar foi baseado em dois critérios: o primeiro sendo as características funcionais e nutricionais dos alimentos; e o segundo, as características dos grupos alimentares com os maiores ou menores *loadings* para essa componente principal, ou seja, os grupos mais importantes para a sua formação.

Tabela 5.3: *Loadings* para os grupos alimentares que verificaram *loadings* elevados ($|> 0.2|$) para as componentes principais extraídas e sujeitas a rotação ortogonal *varimax*[a]

	Comp.1	**Comp.2**	**Comp.3**
Valor próprio	2.1706	2.01185	1.83009
% variância explicada	8.35	7.74	7.04
Vinho	0.3748	-	-
Carnes vermelhas	0.3486	-	-
Bacalhau	0.3308	-	-
Pão branco, integral, tostas, broa	0.2793	-	-
Arroz, massa, batatas cozidas, assadas	0.2505	-	-
Azeite	0.2323	-	-
Cerveja e bebidas brancas	0.2242	-	-
Café	0.2202	-	-
Cereais, bolachas integrais	-0.2128	-	-
Iogurte	-0.3216	-	-
Fruta	-	0.4664	-
Hortícolas	-	0.4505	-
Salada	-	0.4175	-
Peixe (gordo e magro)	-	0.2839	-
Iogurte	-	0.2650	-
Sopa	-	0.2037	-
Snacks	-	-	0.4486
Refrigerantes	-	-	0.4173
Doces	-	-	0.3901
Batatas fritas	-	-	0.3703
Moluscos, crustáceos e peixe de conserva	-	-	0.2101
Peixe (gordo e magro)	-	-	-0.2115

[a] Apenas os grupos alimentares com *loadings* ($|> 0.2|$) para cada componente principal estão presentes na tabela, listados por ordem decrescente para uma interpretação simples e fácil.

O primeiro padrão alimentar identificado reflete uma dieta rural portuguesa, caracterizando-se por um elevado consumo de carnes vermelhas e bacalhau, bastante típicos na região interior, assim como por um consumo elevado de arroz, massa, batatas cozidas e assadas. O azeite também apresentou um consumo elevado, não só sendo usado como condimento (sopa, bacalhau assado, entre outros), como também se encontrando presente na base da maioria dos pratos tradicionais portugueses, que começam por ser preparados a partir de um refogado. O pão branco ou integral, as tostas e a broa apresentaram, como expectável, consumos elevados, uma vez que representam a base da alimentação portuguesa. O vinho apresentou o *loading* mais elevado, sendo este um grande figurante da cultura portuguesa, famosa pela tradição vinícola. Em conjunto com os vinhos, as bebidas brancas e a cerveja apresentaram também *loadings* elevados. Os vinhos e a cerveja são, de maneira geral, as bebidas que acompanham a refeição ou apreciadas em contextos sociais (bares, tascas, cafés ou esplanadas), em conjunto com as bebidas

espirituosas, que servem o propósito de aperitivo ou digestivo. Por conseguinte, o café também reportou um elevado consumo. Este padrão caracterizou-se também por um reduzido consumo de cereais ou bolachas integrais e iogurte. Perante o apresentado, este padrão foi designado por "Dieta tradicional portuguesa".

O segundo padrão foi caracterizado por um elevado consumo de iogurte, peixe e de fruta, hortícolas, salada e sopa. Este padrão é aquele que mais se aproxima da atual noção de uma dieta saudável, partilhando muitas das características com padrões encontrados noutros estudos, frequentemente denominados de *Prudent* ou *Healthy* (Newby and Tucker, 2004). Em Portugal, um país mediterrânico, a ideia de uma dieta saudável ou equilibrada está associada a uma dieta mediterrânica, caracterizada por um elevado consumo de vegetais, frutas, legumes, e cereais, por um consumo moderadamente alto de peixe, por um consumo regular mas moderado de álcool, e um baixo consumo de carnes e produtos láteos (Costacou et al., 2003). De facto, este padrão identificado partilha de algumas características da típica dieta mediterrânica, mas não de todas. A exemplo tem-se o consumo baixo de azeite, contrariamente ao esperado, que poderá encontrar-se substituído por molhos à base de iogurte nas saladas ou outros pratos e/ou consumido com maior moderação e em menores quantidades; o baixo consumo de cereais pouco refinados também não é coerente com a dieta mediterrânica, talvez explicado pelo facto de que a farinha é um componente importante de muitos produtos de pastelaria, alimentos evitados nesta dieta. Outros produtos farináceos, como a massa, arroz e batatas são igualmente pouco consumidos, onde nas refeições se poderão encontrar em baixas porções, compensadas ou complementadas com elevadas porções de hortícolas e salada. A este padrão foi então atribuída a designação "Saudável".

Finalmente, o terceiro padrão alimentar caracterizou-se essencialmente por elevado consumo de *snacks*, refrigerantes, doces, batatas fritas, marisco, e por um baixo consumo de peixe, afastando-se duma dieta tipicamente portuguesa ou que pudesse ser considerada saudável. Este padrão reflete os problemas associados à alimentação frequentemente praticada pelos idosos, uma dieta monótona e insuficiente, derivada de alterações fisiológicas e comportamentais próprias da idade e com consequências na alimentação do idoso. A reduzida vontade de cozinhar e a alteração dos hábitos alimentares por parte de alguns idosos poderá ter origem em alguns fatores sociais ou económicos, como citado na literatura (Herne, 1995; Rosenbloom and Whittington, 1993). Estes estudos reportam, para determinados idosos em certas circunstâncias (por exemplo, que tenham perdido o cônjuge), uma substituição das grandes refeições por refeições simples, rápidas e fáceis de confecionar, como doces, pão, ou outros petiscos/alimentos do seu agrado (Gustafsson and Sidenvall, 2002). Acrescenta-se à decrescente preocupação com a alimentação a reduzida ingestão de água, aparentemente substituída pela ingestão de refrigerantes, como sugerido pelo elevado *loading* para este grupo alimentar. Pelos factos apresentados, foi atribuída a designação "Petiscos" a esta componente principal.

Estas três componentes principais explicaram, respetivamente, 8.35%, 7.74% e 7.04% da variância total do consumo alimentar dos indivíduos da amostra (tabela 5.3), totalizando aproximadamente 23.13% de variância explicada, uma percentagem bastante baixa, mas da mesma ordem de grandeza da apresentada em diversos estudos, onde a formação dos padrões alimentares é feita diretamente a partir das componentes principais extraídas (Chiu et al., 2014; Islam et al., 2014; Kant et al., 1991; Trichopoulos and Lagiou, 2001; Hu et al., 1999; Newby and Tucker, 2004). No entanto, uma vez essa não ser suficientemente razoável para construir diretamente os padrões alimentares, podendo mesmo condicionar a caracterização direta desses (Lopes et al., 2006), optou-se pela realização paralela de uma análise classificatória unicamente com a finalidade de confirmar os padrões obtidos nesta análise (Thorpe et al., 2016; Ocké, 2013).

5.2.2 Análise classificatória

Contrariamente à análise em componentes principais, os métodos usados neste trabalho de análise classificatória permitiram obter um agrupamento dos indivíduos com padrões alimentares semelhantes em categorias mutuamente exclusivas de acordo com o valor das variáveis de consumo alimentar médio diário (Newby and Tucker, 2004). Nesta análise, estas variáveis consistiram nos valores de frequência alimentar média diária dos 26 grupos alimentares, que foram inicialmente centrados e estandardizados de modo a assegurar que todos os grupos tivessem a mesma influência no processo classificatório, dada a sensibilidade desta técnica estatística aos *outliers* e à escala das variáveis. Procedeu-se de seguida à determinação do número ótimo de *clusters* a utilizar com o apoio de critérios de paragem, neste caso o índice Pseudo-F de Calinski-Harabasz, aquando da aplicação do método hierárquico de Ward, seguida da aplicação dos métodos não hierárquicos *K-means* e *K-medians* (soluções a variar de 2 a 8 *clusters*). Na tabela 5.4 encontram-se os índices Pseudo-F de Calinski-Harabasz para as soluções obtidas com os métodos Ward, *K-means* e *K-medians*.

Tabela 5.4: Índices pseudo-F de Calinski-Harabasz para as soluções obtidas com os métodos Ward, *K-means* e *K-medians*

	Índice pseudo-F de Calinski-Harabasz		
Nº Clusters	**Ward**	***K-means***	***K-medians***
2	25.84	62.69	45.52
3	25.88	45.05	36.48
4	27.27	41.68	35.74
5	26.43	39.50	25.44
6	24.97	34.14	24.18
7	24.53	34.90	27.60
8	24.12	35.44	27.40
9	24.63	-	-
10	24.35	-	-
11	23.88	-	-
12	22.81	-	-
13	22.44	-	-
14	22.01	-	-
15	21.45	-	-

Os valores mais elevados dos índices pseudo-F de Calinski-Harabasz obtidos no método de Ward sugeriram a solução de 4 *clusters* para um *clustering* mais distinto ($pseudo - F = 27.27$), no entanto, nos métodos *K-means* e *Kmedians* os índices pseudo-F de Calinski-Harabasz mais elevados foram, por ordem decrescente, para as soluções de 2 e 3 *clusters* ($pseudo - F = 62.69$ e 45.05; 45.52 e 36.48, respetivamente). Note-se que os índices para o método *K-means* são mais elevados do que para o método *K-medians*, sugerindo um *clustering* mais distinto pelo primeiro método.

A interpretabilidade das várias soluções com *clustering* mais distinto para os dois métodos não hierárquicos (isto é, soluções de 2 e 3 *clusters*) e para a solução de 4 *clusters* obtida nesses dois métodos (como sugerido pelo método de Ward) foi analisada. Em ambos os métodos não hierárquicos, concluiu-se que a solução de 2 *clusters* tem pobre interpretabilidade e que a solução de 3 *clusters*, apesar de verificar maior índice pseudo-F, tem menor interpretabilidade que a solução de 4 *clusters*.

A solução de 4 *clusters* do método *K-means* não só apresentou um índice pseudo-F consideravelmente maior do que o da solução obtida com o método *K-medians*, como também uma maior interpretabilidade. Dado o tamanho razoável dos *clusters* formados para um poder estatístico adequado ($> 10\%$ da amostra total) (Thorpe et al., 2016), optou-se por prosseguir o estudo com essa

solução.

Por último, foi avaliada a reprodutibilidade desta solução final, onde a média dos valores das estatísticas Kappa obtidas para as 10 repetições resultantes da aplicação da abordagem de validação cruzada com uma metade aleatória da amostra total foi aproximadamente 0.45, indicando uma concordância moderada das amostras aleatórias com a amostra total. Como tal, esta solução foi considerada uma representação moderadamente fiável dos *clusters* alimentares desta amostra.

5.2.3 Comparação dos padrões obtidos com os dois métodos

As médias e os desvios padrões dos dados centrados e estandardizados do consumo alimentar médio diário para cada grupo alimentar por *cluster* encontram-se representados graficamente na figura 5.2, onde os tons de vermelho representam valores médios positivos de consumo alimentar em determinado grupo, e os tons de azul valores médios negativos de consumo. Note-se que os valores de consumo médio positivos consistem em valores acima da média amostral do consumo médio diário para o respetivo grupo alimentar, e os valores de consumo médio negativos em valores abaixo da média amostral. A intensidade da tonalidade é proporcional à magnitude do valor de consumo médio, permitindo uma comparação de fácil visualização no que toca ao consumo alimentar entre os vários *clusters*, em termos quantitativos e qualitativos.

O uso desta abordagem prendeu-se com o facto de a análise de variância univariada (*Analysis of variance*, ANOVA) por grupo alimentar (e posteriores testes de comparações múltiplas) não poder ser utilizada para comparar os valores de consumo médio, pois não é cumprido o pressuposto de homocedasticidade para todos os grupos alimentares entre os vários *clusters* (teste de Bartlett para igualdade de variâncias com valores-p abaixo de 0.05, rejeitando-se a hipótese nula de homogeneidade de variâncias), recorrendo-se portanto a métodos gráficos.

A descrição dos *clusters*, seguidamente apresentada, foi acompanhada por uma breve discussão que almejou a comparação de cada *cluster* com as componentes principais obtidas anteriormente com o método de análise em componentes principais. Esta comparação teve por base a construção e análise do gráfico da figura 5.3, onde se encontram representadas as médias dos valores dos *scores* das componentes principais por cada *cluster*. Para uma maior facilidade de interpretação, os *scores* respetivos a cada componente principal foram centrados e estandardizados de modo a que os padrões pudessem ser comparados na mesma escala, como sugerido por outros autores (Thorpe et al., 2016).

A designação atribuída a cada padrão alimentar (*cluster*) foi baseada nos mesmos princípios inerentes aos critérios utilizados na análise em componentes principais, em que o primeiro tem em conta as características funcionais e nutricionais dos alimentos. Apesar de não se poder aplicar ANOVA univariada por grupo alimentar (seguida de testes de comparações múltiplas) para concluir ou retirar inferências sobre a igualdade dos respetivos valores médios nos vários *clusters*, estas análises foram realizadas para apoio à decisão de quais *clusters* que se destacam, relativamente aos restantes, quanto ao consumo médio diário de determinado grupo alimentar. Como tal, esse constituiu o segundo critério na caracterização dos grupos alimentares. Note-se que alguns grupos alimentares com uma média mais baixa de valor de consumo alimentar médio diário poderão ser selecionados para representar um *cluster*, enquanto que outros com médias mais elevadas não. A exemplo tem-se a manteiga e o café, onde a manteiga, com uma média de apenas 0.16, foi considerada na caracterização do *Cluster 1*, enquanto que

o café, com uma média de 0.25, não foi considerado, pois esse valor foi baixo relativamente ao *Cluster 4* e elevado relativamente ao *Cluster 2*. Consequentemente, o café foi apenas utilizado na representação do *Cluster 4* e *Cluster 2*.

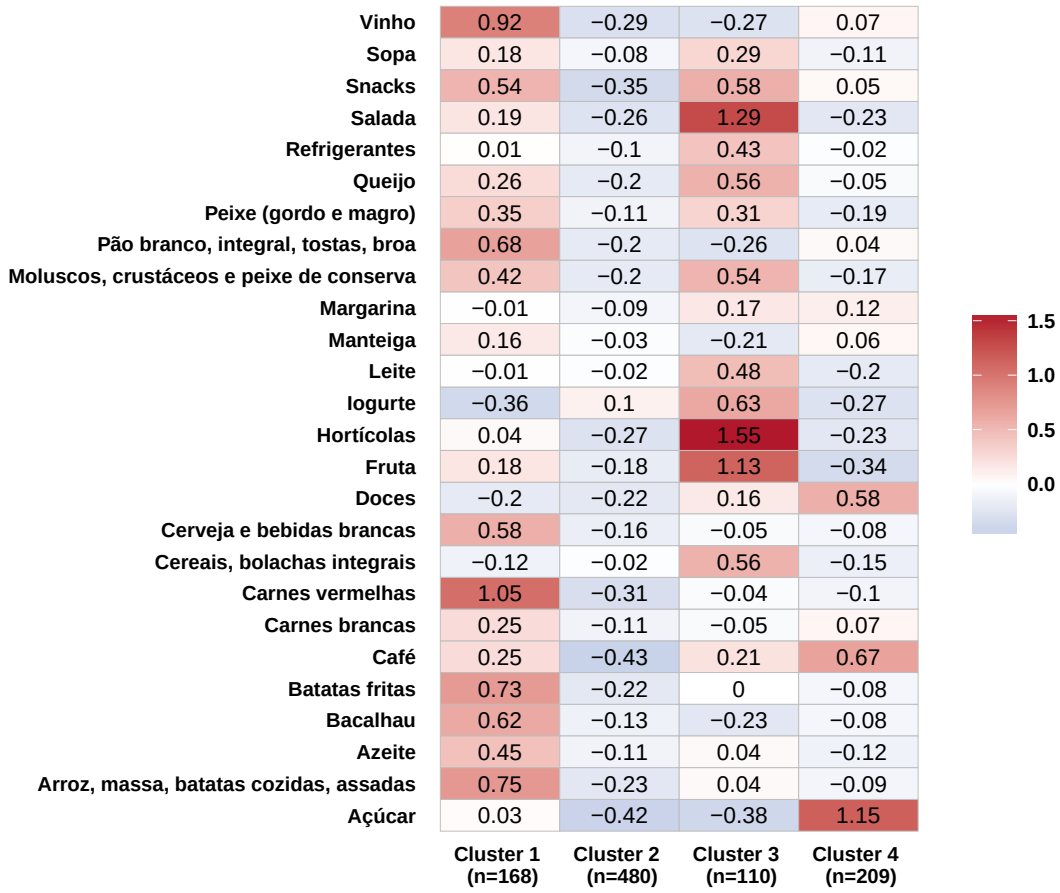


Figura 5.2: Gráfico das médias de consumo médio diário por cada grupo alimentar e por *cluster* (dados centrados e estandardizados; usado o método *K-means*)

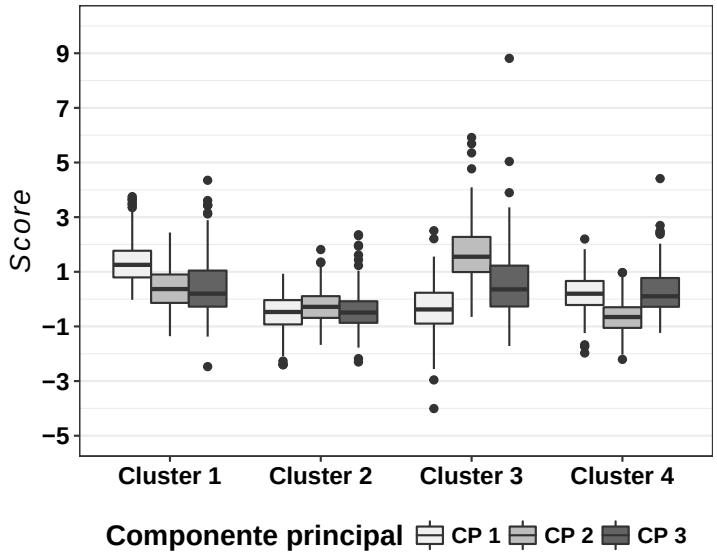
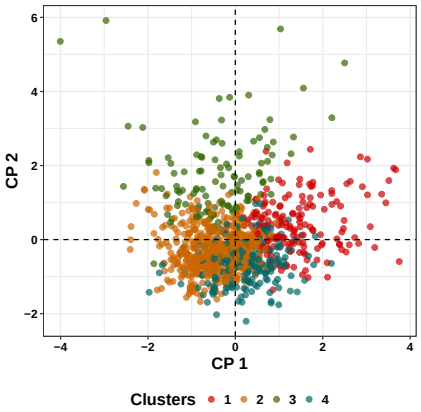
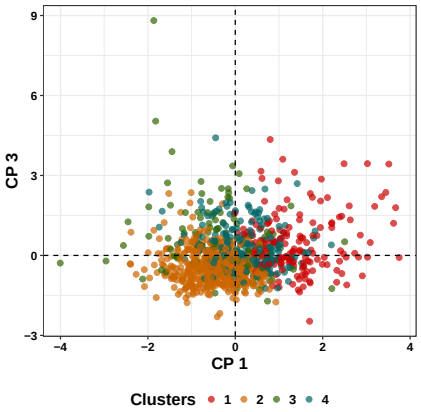


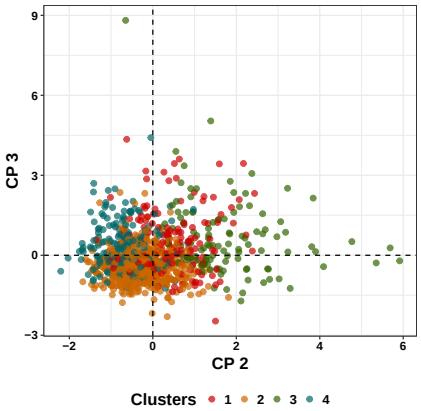
Figura 5.3: Boxplot dos valores dos *scores* obtidos para cada componente principal por *cluster*



(a) Gráfico de *scores* da componente principal 1 (CP 1) contra a componente principal 2 (CP 2)



(b) Gráfico de *scores* da componente principal 1 (CP 1) contra a componente principal 3 (CP 3)



(c) Gráfico de *scores* da componente principal 2 (CP 2) contra a componente principal 3 (CP 3)

Figura 5.4: Gráficos de *scores* das componentes principais, coloridos por *cluster*

• ***Cluster* 1**

O primeiro padrão identificado ($n = 168$) caracterizou-se essencialmente por um elevado consumo de carnes vermelhas, vinho, acompanhamentos, batatas fritas, pão, bacalhau, cerveja e bebidas brancas e *snacks*; por um consumo moderado a elevado de azeite, moluscos, peixe, queijo e carnes brancas; e por um consumo moderado de manteiga. Verificou-se também um baixo consumo de iogurte, doces e cereais. Como se pode ver, este *cluster* parece partilhar muitas características com a componente principal 1 (carnes vermelhas, bacalhau, pão ou tostas, acompanhamentos, azeite, bebidas alcoólicas e café). Tal foi corroborado, efetivamente, pela análise do gráfico da figura 5.3, onde indivíduos deste *cluster* apresentaram *scores* mais elevados na componente principal 1 (média: 1.40, IC 95%: [1.29, 1.53]), seguida pela componente 3 (*snacks*, refrigerantes, doces, batatas fritas e marisco) (0.46, [0.29, 0.64]) e pela componente 2 (fruta, hortícolas, salada, peixe, iogurte e sopa) (0.41, [0.30, 0.52]). Pelas razões apresentadas, a este *cluster* atribuiu-se a designação de "Dieta tradicional portuguesa".

• ***Cluster* 2**

O segundo padrão identificado ($n = 480$) corresponde ao maior *cluster* identificado e apresentou

uma baixa média de consumo para uma grande parte dos grupos alimentares comparativamente aos outros três *clusters*. Este padrão apresenta consistência com o argumento anteriormente apresentado, que sugere uma alimentação pobre e insuficiente por parte de alguns idosos sob determinadas condições de saúde ou determinadas condições sociais ou económicas (Herne, 1995; Rosenbloom and Whittington, 1993). Consequentemente, os indivíduos pertencentes a este *cluster* demonstraram um baixo *score* para todas as componentes principais (componente 1: -0.48, $[-0.53, -0.43]$; componente 2: -0.27, $[-0.33, -0.22]$; componente 3: -0.44, $[-0.49, -0.38]$).

Este *cluster* encontra-se frequentemente reportado na literatura (Thorpe et al., 2016) com a designação de "Small eaters", aqui também adoptada, não demonstrando nenhuma tendência ou dominância de grupos alimentares. Nenhum padrão equivalente foi identificado pela análise em componentes principais, talvez pelo facto de esta técnica se basear nas correlações entre as variáveis iniciais, em vez de nos valores absolutos, como a análise classificatória.

- ***Cluster* 3**

O terceiro padrão ($n = 110$) foi caracterizado por um elevado consumo de hortícolas, salada, fruta, produtos lácteos (iogurte, queijo e leite), cereais e moluscos. Apresentou também um consumo moderado de peixe, sopa, refrigerantes e de doces. Os indivíduos deste *cluster* apresentaram uma baixa ingestão de manteiga, açúcar, bebidas alcoólicas e pão. Este *cluster* parece aproximar-se da componente principal 2 designada de "Saudável", no entanto, o elevado consumo de refrigerantes e *snacks* e o consumo moderado de doces aproxima-o simultaneamente à componente principal 3 "Petiscos", o que não era esperado. Tal pode ocorrer pelo facto de o grupo alimentar dos refrigerantes abranger diversos tipos de produtos, não só os produtos gaseificados, mas também néctares, sumos de fruta embalados, entre outros; assim como o grupo alimentar dos *snacks*, que inclui enchidos, salsichas, ovos, salgados, hambúrguer e *pizza*. O elevado consumo de refrigerantes poderá estar maioritariamente associado ao consumo de néctares ou sumos de fruta embalados ou naturais, frequentemente considerados como saudáveis. De facto, ao analisar separadamente a distruibuição do consumo médio diário para cada constituinte (item do questionário) da variável `refriger`, verificou-se uma maior de frequência para o item alimentar que contém os sumos naturais e néctares (ver figura 7.4, Apêndice C), no entanto, esse item alimentar também contém outros refrigerantes (ver questionário alimentar, Apêndice A). Da mesma maneira, dentro do grupo dos *snacks* está incluído o ovo, que é não só considerado um alimento benéfico para a saúde e muito nutritivo, como também é um alimento de fácil preparação e de baixo custo. O elevado consumo de *snacks* e o consumo moderado de doces sugerem também que, apesar de os indivíduos praticarem de maneira geral uma alimentação saudável, não se privam de determinados alimentos ao seu gosto, como os enchidos, doces ou outros. Este *cluster* foi consequentemente designado de "Dieta equilibrada".

Como esperado, os indivíduos deste *cluster* apresentaram elevados *scores* para a componente 2 (1.79, $[1.58, 2.01]$), seguida pela componente 3 (0.63, $[0.36, 0.90]$) e baixos *scores* para a componente 1 (-0.41, $[-0.60, -0.22]$).

- ***Cluster* 4**

Por último, o quarto padrão identificado ($n = 209$) apresentou um elevado consumo de açúcar, café e doces, e um baixo consumo de iogurte, queijo, sopa, peixe, moluscos, hortícola, salada e fruta. Mais uma vez, estes resultados poderão refletir alguns dos problemas associados à má alimentação

praticada por alguns idosos sob determinadas condições, onde parece existir uma tendência de substituir as principais refeições por doces ou merendas fáceis de preparar, como por exemplo pão com manteiga ou enchidos (pois os *snacks* apresentam valores positivos médios). O elevado consumo de café e açúcar corrobora esta afirmação, dado que o consumo dos alimentos acima referido é usualmente acompanhado por café ou o tradicional chá da tarde.

Os indivíduos deste *cluster* verificaram, como esperado, baixos *scores* para a componente 1 e 3 (componente 1: 0.19, $[0.10,\ 0.29]$; componente 3: 0.30, $[0.18,\ 0.42]$) e *scores* negativos para a componente 2 (-0.65, $[-0.72,\ -0.57]$). Os baixos *scores* evidenciam também uma alimentação insuficiente, pelo que a este *cluster* foi atribuída a designação "Petiscos".

A observação dos gráficos da figura 5.4 corrobora com as descrições acima apresentadas. Parecem haver mais indivíduos do *Cluster* 1 que apresentaram valores de *scores* para as componentes principais 2 e 3 (CP 2 e CP 3) acima da média, com grande dispersão, apesar de ainda bastantes participantes terem apresentado valores de *score* abaixo da média. Todos os indivíduos do *Cluster* 1, à excepção de apenas um participante, apresentaram valores de *scores* para a componente principal 1 (CP1) acima da média.

A maioria dos indivíduos do *Cluster* 2 tiveram valores de *scores* para a componente principal 1 abaixo da média. Para as componentes principais 2 e 3 existiram ainda bastantes participantes acima da média, no entanto, grande parte dos indivíduos teve valores abaixo da média. Note-se que os valores de *scores* acima da média foram, de modo geral, valores baixos.

À exceção de um indivíduo, todos os partipantes do *Cluster* 3 verificaram valores de *scores* para a componente principal 2 acima da média, existindo grande dispersão dos valores positivos. Para a maioria dos indivíduos, os *scores* para a componente principal 3 foi acima da média, onde os valores positivos apresentaram grande dispersão, enquanto que para a componente principal, a maioria verificou *scores* negativos. Os valores absolutos dos *scores* para esta componente apresentaram grande dispersão.

Finalmente, a maior parte dos indivíduos do *Cluster* 4 verificou valores de *scores* para a componente principal 3 acima da média; para a componente principal 1, maior proporção de indivíduos apresentou *scores* acima da média, mas ainda assim muitos apresentaram *scores* abaixo da média. Para a componente principal 2, poucos indivíduos apresentaram *scores* acima da média.

Perante a descrição dos *clusters* aqui apresentada e respetiva comparação com as diferentes componentes principais obtidas, pode concluir-se que existe consistência nos padrões alimentares identificados com as duas técnicas estatísticas, apesar da presença de algumas diferenças nos resultados obtidos por cada uma. Foram então utilizados os padrões extraídos com a análise em componente principais nas análises seguintes.

5.2.4 Características dos indivíduos por padrão alimentar

De modo a compreender e estudar a distribuição destes padrões na população em estudo, foram calculados os *scores* de cada indivíduo para cada uma das componentes principais extraídas. Estes foram divididos em tercis: o primeiro tercil incluiu indivíduos com baixa adesão ao padrão alimentar; o segundo tercil uma adesão moderada e o terceiro tercil uma adesão elevada ao padrão alimentar. Na tabela 7.11 do Apêndice C encontram-se estatísticas descritivas dos *scores* para as componentes

principais (mediana, amplitude interquartil, mínimo, máximo e valores do 1º e 3º tercil) e nas figuras 7.5 e 7.6 do Apêndice C encontra-se um gráfico boxplot dos *scores* para cada componente principal e gráficos boxplot da distribuição da idade por tercil de cada componente principal, respetivamente.

Sendo a idade um fator de risco para o desenvolvimento e progressão da DMI, considerou-se importante verificar inicialmente, para cada componente principal analisada, se a distribuição etária entre os vários tercis de cada componente principal era homogénea. Com esse objetivo, foi aplicado o teste de qui-quadrado de homogeneidade após a categorização da variável idade em três níveis: "55 - 64 anos", "65 - 74 anos", "$\geq$ 75 anos". Estas categorias foram sugeridas em *Prevalence of Age-Related Macular Degeneration in Portugal: The Coimbra Eye Study - Report 1* (Cachulo et al., 2015), sendo que as categorias "75 - 84 anos"e "$\geq$ 85 anos"utilizadas nesse artigo foram aqui agrupadas devido ao reduzido número de participantes na última faixa etária quando divididos pelos vários tercis de cada componente principal, podendo originar problemas subsequentes na realização do teste de qui-quadrado.

Na tabela 7.12 do Apêndice C encontram-se as frequências de indivíduos por cada faixa etária e por tercis de cada componente principal. Encontram-se também as proporções de indivíduos em cada grupo de idades dentro de cada tercil de determinada componente principal e os valores-p dos testes de homogeneidade de qui-quadrado acima referidos. Dado rejeitar-se a hipótese nula de homogeneidade da distribuição etária dos tercis nas três componentes principais (valor-$p < 0.001$), procedeu-se à estandardização por idade de todas as variáveis (à excepção da variável idade) referentes às características sociais, económicas, demográficas e respeitantes ao estilo de vida e historial médico avaliadas (procedimento completo descrito no Apêndice C) (Chiu et al., 2014).

Na figura 5.5 encontram-se gráficos dos valores estandardizados para a idade da média (variáveis contínuas) e das proporções (variáveis categóricas) por tercil e por cada componente principal de todas as características a ser analisadas. Com estes gráficos podemos obter uma ideia geral do comportamento das variáveis não só entre os tercis de cada componente principal, mas também entre as várias componentes.

A observação do gráfico da idade evidencia um decréscimo da adesão a cada componente principal com o aumento da idade. Estas diferenças etárias foram mínimas entre os vários tercis de cada componente principal, assim como entre as várias componentes principais.

Os indivíduos que têm menor adesão ao padrão "Dieta Tradicional Portuguesa"parecem ser em menor proporção do sexo masculino (aproximadamente 0.25), sendo que esta verifica um considerável aumento com a maior adesão a esse padrão, chegando mesmo a atingir uma proporção próxima de 0.75. Pelo contrário, o padrão "Saudável"é aquele que apresenta proporções mais constantes ao longo dos tercis, onde a proporção de homens aparenta ser ligeiramente inferior à das mulheres. Quanto ao padrão "Petiscos", parece existir uma maior proporção de mulheres no primeiro tercil, verificando-se um aumento ligeiro na proporção de homens com o aumento da adesão ao padrão, quase igualando a proporção de ambos os sexos neste tercil.

Segundo a Direção-Geral da Saúde (DGS), os valores de referência de consumo energético e os valores de referência para o perímetro abdominal são superiores para os homens (quando comparados com os das mulheres) (Direção-Geral da Saúde, 2005). Como tal, optou-se por se fazer a análise estratificada por sexo das variáveis `cia_energia_kcal`, `perimetroabd` e `imc`, tendo sido construídos gráficos separados para cada uma das variáveis e para cada género.

Entre as mulheres, os padrões "Saudável"e "Petiscos"foram aqueles que verificaram um maior aumento dos valores do perímetro abdominal. O aumento do perímetro abdominal ao longo dos tercis do padrão "Petiscos"aparentou ser linear e de aproximadamente 5 cm. Já para o padrão "Saudável"o aumento observado ao longo dos tercis foi de menor magnitude (cerca de 3 cm) e mais acentuado do tercil adesão moderada para o tercil de adesão elevada. O padrão "Dieta tradicional portuguesa"foi aquele que verificou um maior valor de perímetro abdominal no tercil de adesão baixa, que sofreu uma diminuição no segundo tercil e novamente um aumento no terceiro tercil. Quanto aos valores de IMC, verificou-se um aumento aproximadamente linear de cerca de 0.61 $\mathrm{kg/m^2}$ ao longo dos tercis do padrão "Saudável"e um aumento de cerca de 0.73 $\mathrm{kg/m^2}$ ao longo dos tercis do padrão "Petiscos", sendo esse menos acentuado do segundo para o terceiro tercil. Relativamente ao padrão "Dieta tradicional", este foi aquele que mais uma vez verificou maior valor de IMC médio no primeiro tercil, que não se alterou significativamente no segundo tercil, mas sofreu um aumento acentuado de cerca de 0.80 $\mathrm{kg/m^2}$ no terceiro tercil. O consumo energético médio diário aparentou aumentar de forma similar ao longo dos tercis de todos os padrões, à excepção do aumento do segundo para o terceiro tercil do padrão "Dieta tradicional portuguesa", que pareceu ser de maior magnitude. Além disso, os valores de consumo médio energético diário para este padrão pareceram, de modo geral, ser superiores aos dos outros padrões ao longo dos tercis.

Entre os homens, o padrão "Dieta tradicional portuguesa"foi aquele que apresentou um maior aumento do perímetro abdominal médio ao longo dos tercis, sendo esse valor próximo de 6 cm. Este padrão foi também aquele que apresentou menor valor no primeiro tercil (99.43 cm). Quanto ao padrão "Saudável", este foi o que apresentou maior valor de perímetro abdominal médio no primeiro tercil, que diminuiu no segundo tercil apenas cerca de 1.6 cm e aparentemente se manteve no terceiro tercil. Finalmente, no padrão "Petiscos"observou-se um aumento de aproximadamente 3 cm do valor do perímetro abdominal médio, sendo esse mais substancial do segundo para o terceiro tercil. Quanto aos valores de IMC médios, ao longo dos tercis verificou-se igualmente um grande aumento para o padrão "Dieta tradicional portuguesa"(cerca de 1.45 $\mathrm{kg/m^2}$), um aumento próximo de 0.92 $\mathrm{kg/m^2}$ no padrão "Petiscos"e uma ligeira diminuição no padrão "Saudável"(0.3 $\mathrm{kg/m^2}$). Mais uma vez, o padrão "Dieta tradicional portuguesa"foi aquele que apresentou menor valor de IMC médio no primeiro tercil e o padrão "Saudável"o maior valor. O consumo energético médio diário aumentou ao longo dos tercis de todos os padrões de forma similar, embora de maneira geral os valores do padrão "Dieta tradicional portuguesa"tenham sido ligeiramente inferiores, e os valores do padrão "Saudável"tenham sido ligeiramente superiores nos tercis de consumo moderado e elevado.

É de salientar que uma grande proporção de participantes do sexo feminino e masculino apresentam IMC e perímetro abdominal acima dos valores de referência. Segundo a DGS, as mulheres de perímetro abdominal igual ou superior a 80 cm apresentam risco metabólico aumentado e acima de 88 cm de risco metabólico muito aumentado. Já para os homens estes valores são de 94 cm e 104 cm, respetivamente (Direção-Geral da Saúde, 2005). Quanto ao IMC, os participantes (de ambos os sexos) com valores acima de 24.9 $\mathrm{kg/m^2}$ apresentam risco metabólico aumentado (Direção-Geral da Saúde, 2005). 70.71% das mulheres apresentaram um perímetro abdominal igual ou superior a 88 cm e 19.02% das mulheres um perímetro abdominal igual ou superior a 80 cm e inferior a 88 cm; e nos homens estas percentagens foram de 48.72% e 34.57%, respetivamente.

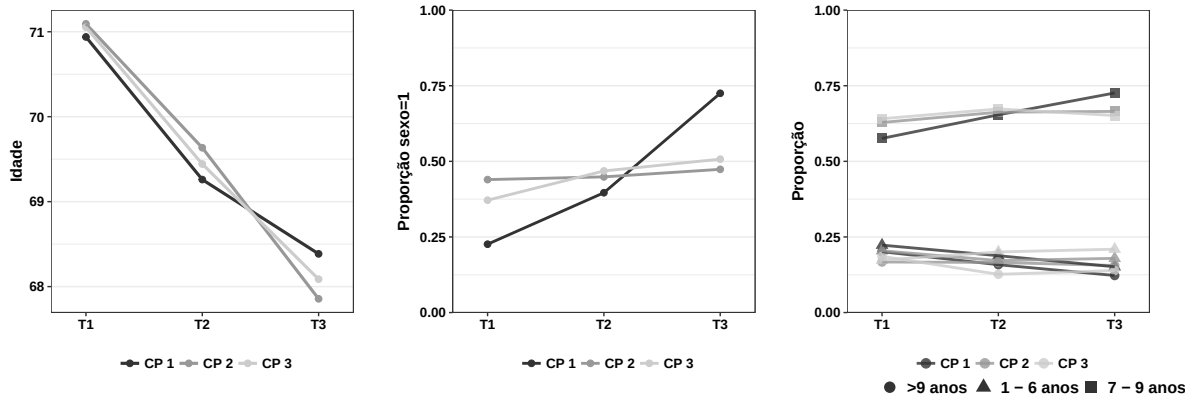
A proporção de indivíduos diabéticos, hipertensos e com dislipidemias revelou também ser similar nos três padrões principais, mantendo-se aproximadamente constante ao longo dos tercis. Quanto à

proporção de diabéticos, este valor andou à volta de 0.25; já a proporção de hipertensos aparentou ser bastante superior, andando próxima de 0.70; a proporção de dislipidémicos no primeiro tercil do padrão "Petiscos"aparentou ser menor comparativamente aos outros padrões, aumentando ao longo dos tercis e ultrapassando os valores verificados para os outros padrões, onde a proporção de dislipidémicos diminuiu ligeiramente do primeiro tercil para o segundo tercil. Estas proporções variaram entre os valores de 0.51 e 0.61, não verificando, assim, grandes alterações.

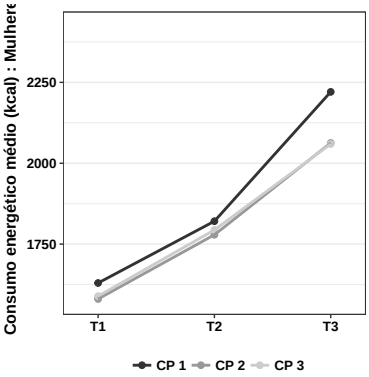
A proporção de indivíduos praticantes de exercício físico regular pareceu ser semelhante e aproximadamente constante nos três padrões, rondando o valor de 0.25. Note-se o ligeiro aumento no padrão "Saudável"com o aumento à sua adesão.

Com o aumento da adesão ao padrão "Dieta Tradicional Portuguesa"pareceu existir um aumento considerável na proporção de fumadores ou ex-fumadores, acompanhado por um aumento acentuado do número de pacotes de tabaco anuais consumidos e por um aumento substancial do consumo álcoolico. Os restantes padrões apresentam comportamentos semelhantes entre si, onde as proporções de fumadores ou ex-fumadores e os valores de consumo alcoólico médio não pareceram diferir significativamente ao longo dos tercis, apresentando valores próximos de 0.25 e 10 a 11g/dia, respetivamente. Foi apenas reportado um aumento ligeiro do consumo alcoólico médio do primeiro para o segundo tercil no padrão "Petiscos". Relativamente ao número de pacotes de tabaco consumidos anualmente, este aparentou sofrer um aumento ao longo dos tercis do padrão "Petiscos", contrariamente ao observado para o padrão "Saudável", onde se observou um aumento de pacotes do primeiro para o segundo tercil, seguido de uma diminuição para o terceiro tercil.

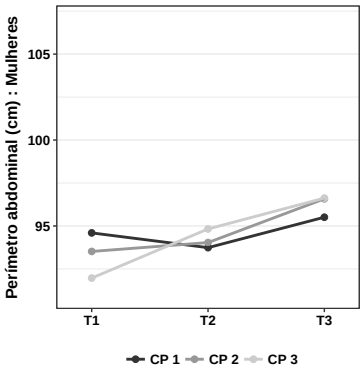
Em todas as componentes principais, a percentagem de indivíduos com 1 a 6 anos de escolaridade rondou os 60%, diferindo consideravelmente das percentagens dos indivíduos com diferentes anos de escolaridade (abaixo dos 25%). Estas proporções pareceram ser aproximadamente constantes ao longo dos tercis de cada componente principal, apesar de no padrão "Dieta Tradicional Portuguesa"parecer existir um ligeiro aumento na proporção de indivíduos com 1 a 6 anos de escolaridade.



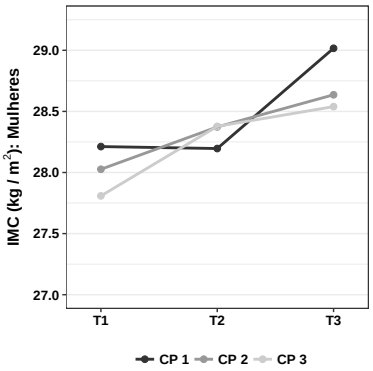
(a) Idade (b) Sexo (c) Escolaridade



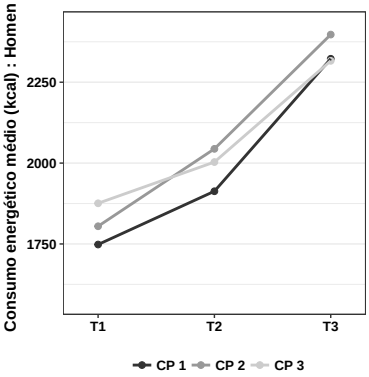
(d) Consumo energético médio diário: mulheres



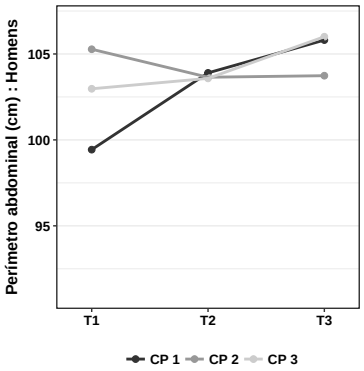
(e) Perímetro abdominal: mulheres



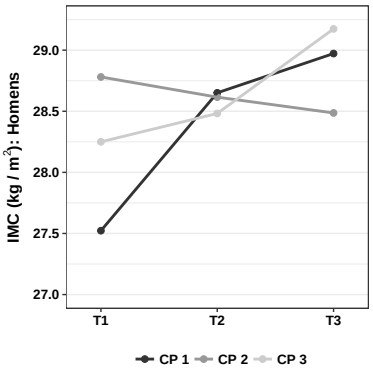
(f) Índice de massa corporal: mulheres



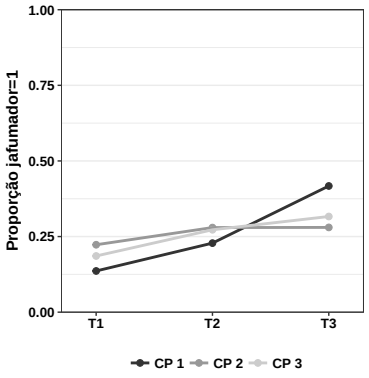
(g) Consumo energético médio diário: homens



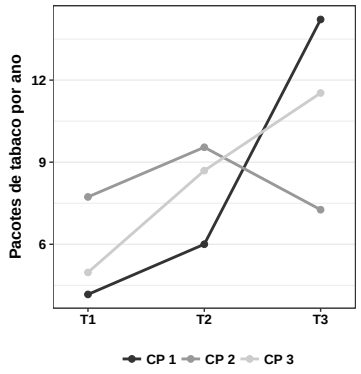
(h) Perímetro abdominal: homens



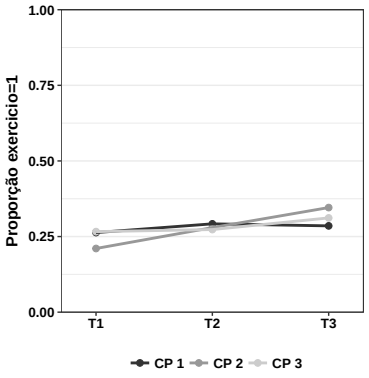
(i) Índice de massa corporal: homens



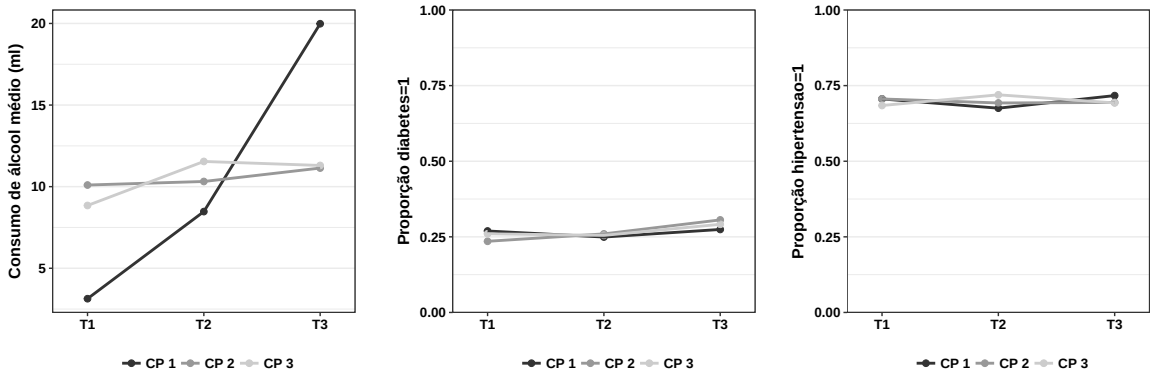
(j) Hábitos tabágicos



(k) Pacotes de tabaco por ano



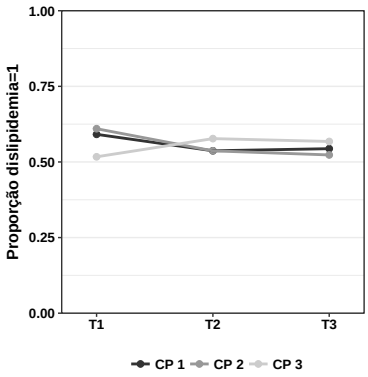
(l) Exercício físico regular



(m) Consumo de álcool

(n) Diabetes

(o) Hipertensão



(p) Dislipidémia

Figura 5.5: Gráficos das proporções (variáveis categóricas) e das médias (variáveis contínuas) das características demográficas por tercil (T) de cada componente principal estandardizada para a idade (CP)

5.2.5 Composição nutricional por padrão alimentar

Apesar de os padrões alimentares terem sido construídos apenas com base em informação relativa aos grupos alimentares, a ingestão de nutrientes constitui também um bom elemento de caracterização dos padrões, como sugerido por (Lopes et al., 2006). Uma vez que maiores valores absolutos de consumo energético tendem a resultar em maiores consumos de nutrientes, os nutrientes foram submetidos a um ajustamento por energia, que consiste num método analítico pelo qual o consumo de cada nutriente é avaliado em relação ao consumo de energia total, aplicando o método residual proposto por Willett e Stampfer (Willett et al., 1997). Este método consiste na construção de um modelo de regressão linear, onde a variável dependente é o consumo absoluto diário do nutriente a ser ajustado e a variável independente o consumo energético médio diário (variável `cia_energia_kcal`). O consumo dos nutrientes ajustados para energia são assim calculados como os resíduos desse modelo de regressão. Adicionalmente, foram estandardizados através da sua divisão pelo respetivo desvio padrão de modo a permitir maior comparabilidade entre as diferentes componentes principais.

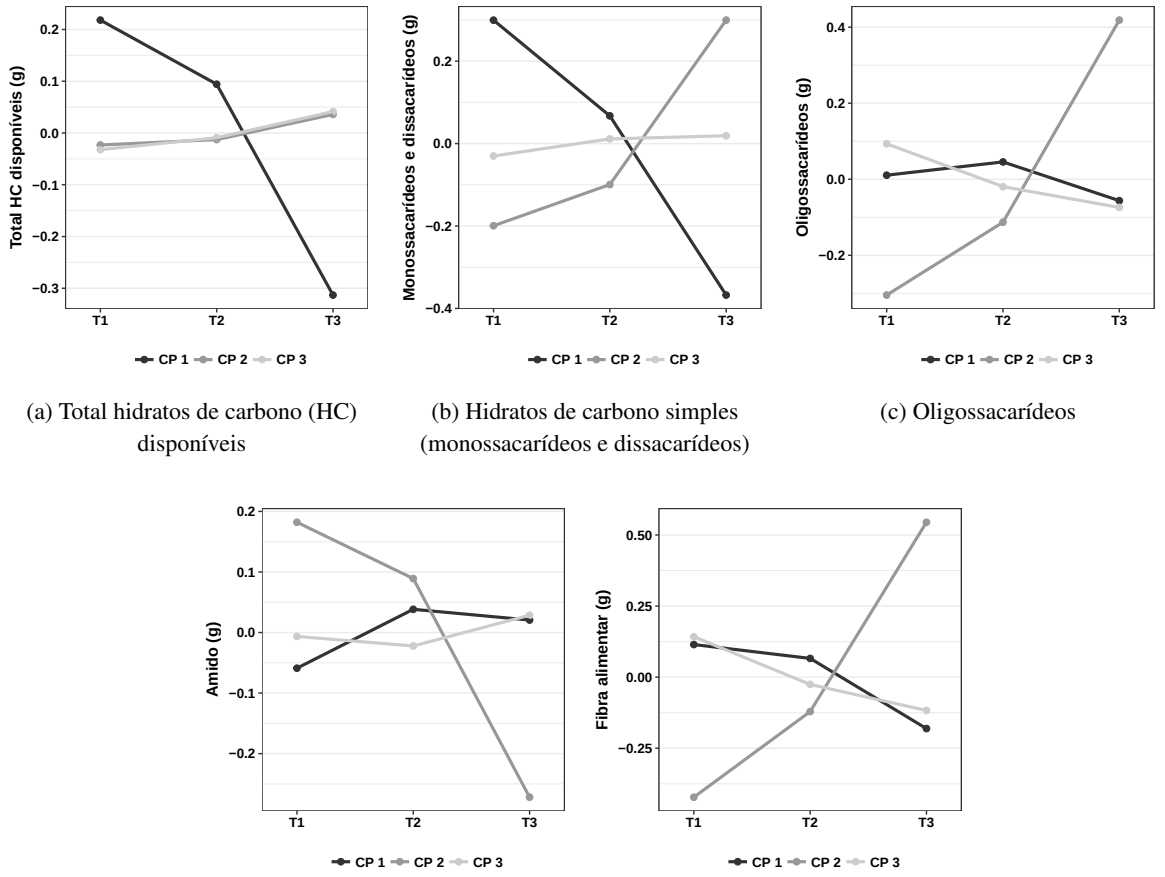
A composição nutricional de cada alimento foi calculada com base na Tabela de Composição de Alimentos Portuguesa (INSA, 2006), onde na tabela 7.2 do Apêndice B se encontram os valores nutricionais para cada item alimentar.

Seguidamente serão apresentados gráficos das médias dos hidratos de carbono, gorduras, antioxidantes e vitaminas (ajustados para energia) por cada tercil e componente principal. Será também

feita uma breve descrição e discussão dos resultados obtidos, tendo em conta os valores nutricionais da tabela 7.2. Note-se que os nutrientes utilizados na análise aqui apresentada foram previamente selecionados, tendo sido primeiramente realizada uma breve revisão bibliográfica sobre a associação de todos os nutrientes disponíveis na base de dados com o risco de DMI e escolhidos aqueles que poderiam, potencialmente, influenciar o seu desenvolvimento e/ou progressão.

- **Hidratos de carbono**

Na figura 5.6 encontram-se representados os gráficos das médias do consumo alimentar médio diário por tercil de cada componente principal dos vários hidratos de carbono disponíveis na base de dados original e selecionados neste estudo para sequente avaliação. Estes nutrientes encontram-se ajustados por energia.



(a) Total hidratos de carbono (HC) disponíveis

(b) Hidratos de carbono simples (monossacarídeos e dissacarídeos)

(c) Oligossacarídeos

(d) Amido

(e) Fibra alimentar

Figura 5.6: Gráficos das médias dos hidratos de carbono (ajustados para energia) no tercil (T) de cada componente principal (CP)

A análise do gráfico do consumo total de hidratos de carbono sugere que o consumo nos padrões "Saudável"e "Petiscos"aumenta ligeiramente ao longo dos tercis dessas componentes, ao contrário do que se verificou para o padrão "Dieta Tradicional Portuguesa". Neste padrão, o consumo total médio de hidratos de carbono disponíveis apresentou um valor muito elevado no primeiro tercil, que diminuiu

significativamente ao longo dos restantes tercis.

O padrão "Saudável"aparentou ser aquele com maiores valores de consumo médio diário de hidratos de carbono simples, oligossacarídeos e fibra alimentar no terceiro tercil. Os monossacarídeos e os dissacarídeos são maioritariamente provenientes de fontes alimentares que incluem o açúcar e produtos açúcarados como o mel, doces, alguns refrigerantes e, em menor quantidade, o leite e iogurte, fruta e alguns vegetais. Relativamente aos hidratos de carbono complexos, os oligossacarídeos e a fibra alimentar estão presentes em baixas quantidades nos itens alimentares, embora em maior quantidade nos gelados de leite, cebola, alguns hortícolas (cebola, ervilhas, brócolos), na sopa e na carne de perú ou coelho no caso dos oligossacarídeos; e nas batatas fritas de pacotes, pão ou tostas integrais, chocolate e alguns hortícolas (leguminosas, ervilhas), no caso da fibra alimentar. Uma vez que o padrão "Saudável"se caracterizou por um elevado consumo de iogurte, de fruta, hortícolas, salada e sopa, justifica-se o aumento desses hidratos de carbono com o aumento da adesão a este padrão.

O padrão "Dieta Tradicional Portuguesa"apresentou uma acentuada diminuição no consumo de hidratos de carbono simples ao longo dos tercis e um consumo aproximadamente constante para os restantes nutrientes acima descritos, sofrendo uma ligeira diminuição para a fibra alimentar. É de facto expectável que os valores dos hidratos de carbono simples diminuam bastante com o aumento dos tercis, uma vez que este padrão se caracteriza por um baixo consumo de produtos lácteos, e onde o consumo de arroz, massa, batatas e pão ou tostas têm especial destaque (ao contrário da fruta e hortícolas). Estes alimentos constituem também das principais fontes alimentares do amido, o que poderá explicar o ligeiro aumento dos respetivos valores de consumo médio com a maior adesão a este padrão.

O padrão "Petiscos"foi aquele que aparentou ser mais estável no consumo ao longo dos tercis, apresentando valores de consumo médio próximos da média.

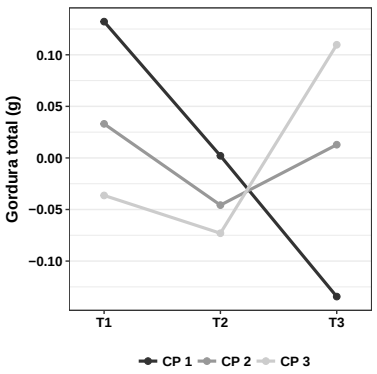
- **Gorduras**

Na figura 5.7 encontram-se representados os gráficos das médias do consumo alimentar médio diário por tercil de cada componente principal das várias gorduras disponíveis na base de dados original e selecionadas neste estudo para sequente avaliação. Estes nutrientes encontram-se ajustados por energia.

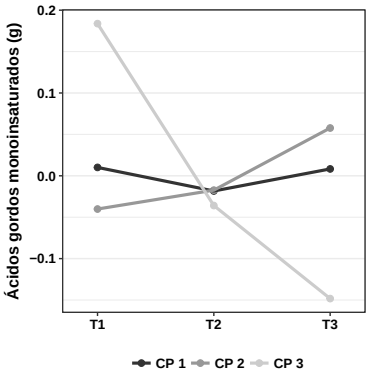
O padrão "Petiscos"foi aquele que apresentou valores de consumo médio mais elevados de ácidos gordos *trans*, ácidos gordos saturados, colesterol e ácidos gordos polinsaturados (entre estes últimos encontra-se também o ácido linoleico). Os ácidos gordos *trans* e saturados são abundantes em alimentos como margarina, manteiga, produtos de confeitaria e pastelaria, alguns gelados, *snacks* (fiambre, enchidos, ovos, pizza, hambúrguer, salgados), bolachas, batatas fritas e carnes vermelhas, alimentos estes que refletem uma grande maioria dos constituintes com consumo elevado no padrão "Petiscos". Muitos desses alimentos são também ricos em colesterol, onde os ovos são o alimento com maior valor de colesterol por 100g relativamente aos restantes, seguido pelas lulas, polvo, camarão, mexilhão, entre outros moluscos, também bastante consumidos neste padrão. Deste modo, o aumento do consumo médio dos ácidos gordos *trans*, saturados e colesterol ao longo dos tercis do padrão "Petiscos"é facilmente compreendido.

Os ácidos gordos polinsaturados estão presentes em grande quantidade nos óleos vegetais (girassol, milho, soja) e, embora em menor quantidade, nos frutos oleoginosos (avelãs, amendoins, amêndoas, nozes), margarina, peixe de conserva, batatas fritas, azeite, bacon ou toucinho, no peixe gordo (gordura

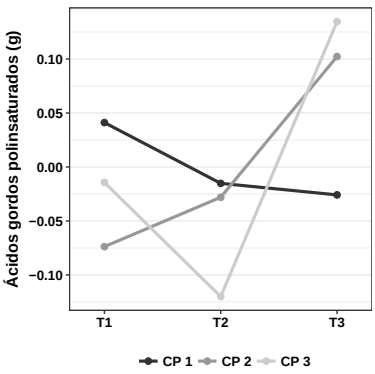
e óleo de peixe), croquetes, produtos de pastelaria ou confeitaria e salgados. Os ácidos gordos polinsaturados incluem dois tipos principais de ácidos gordos essenciais não produzidos pelo organismo humano, apenas obtidos pela alimentação, os ácidos gordos ómega-3 e ómega-6. O ácido linoleico é um ácido gordo essencial da família ómega-6 muito encontrado nos óleos vegetais. Desse modo, os alimentos que utilizam óleo na sua confeção, como as batatas fritas, salgados ou produtos de pastelaria maionese, contêm um elevado teor de ácido linoleico. Estes alimentos apresentam um elevado consumo no padrão "Petiscos", explicando tanto o aumento acentuado desse ácido gordo do tercil de consumo moderado para o tercil de consumo elevado, como também o aumento dos ácidos gordos polinsaturados.



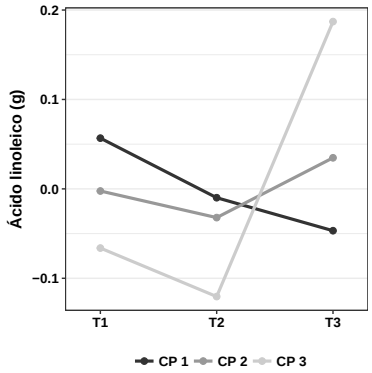
(a) Gordura total



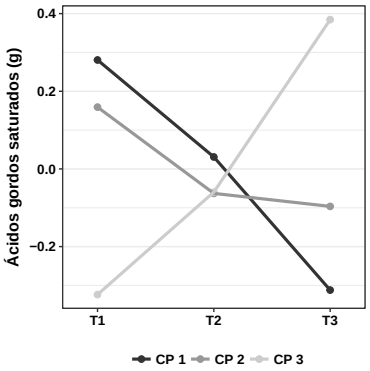
(b) Ácidos gordos monoinsaturados



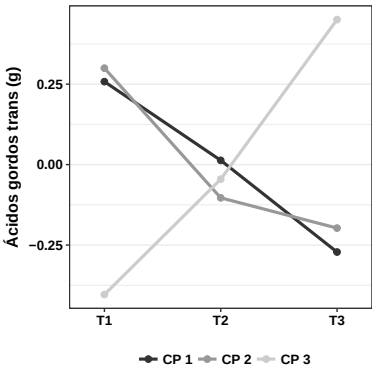
(c) Ácidos gordos polinsaturados



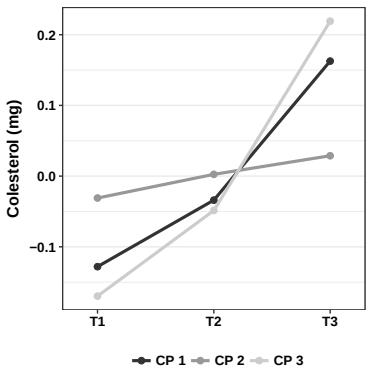
(d) Ácido linoleico (ácido gordo polinsaturado ómega-6)



(e) Ácidos gordos trans



(f) Ácidos gordos saturados



(g) Colesterol

Figura 5.7: Gráficos das médias das gorduras (nutrientes ajustados para energia) no tercil (T) de cada componente principal (CP)

É curioso verificar que no padrão "Saudável"existiu, de igual modo, um aumento acentuado ao longo dos tercis de ácidos gordos polinsaturados, não se verificando, no entanto, o mesmo para o consumo médio de ácido linoleico, que permaneceu aproximadamente constante. Note-se que os ácidos gordos polinsaturados são de diversos tipos, e aqui apenas se teve registo para o consumo particular do ácido linoleico. Pode então deduzir-se que no padrão "Saudável"os ácidos gordos poderão provir essencialmente de outras fontes alimentares mais saudáveis ricas noutros tipos de ácidos polinsaturados (Seddon et al., 2003b). A exemplo tem-se o peixe, que apresenta níveis elevados de ácidos gordos ómega-3, e os hortícolas. O consumo de peixe, hortícolas, saladas e sopa encontra-se também normalmente associado ao consumo de azeite, a fonte principal de ácidos gordos monoinsaturados, uma vez que é usado como condimento e/ou no processo da sua preparação. Tal poderá estar na base do ligeiro aumento do consumo médio de ácidos gordos monoinsaturados observado ao longo dos tercis.

Relativamente aos ácidos gordos *trans* e gorduras saturadas, o padrão "Saudável"apresentou valores elevados de consumo médio no primeiro tercil, que diminuíram com o aumento da adesão ao padrão. Salienta-se também que este padrão foi aquele que apresentou níveis médios de colesterol mais estáveis ao longo dos tercis, apenas com um ligeiro aumento.

O padrão "Dieta tradicional portuguesa"verificou igualmente uma diminuição acentuada dos ácidos gordos *trans* e saturados ao longo dos tercis. Os restantes nutrientes apresentaram valores de consumo médio aproximadamente constantes, onde se reportou apenas uma ligeira diminuição dos níveis médios de ácidos gordos polinsaturados e ácido linoleico ao longo dos tercis.

• **Antioxidantes, sais minerais e proteínas**

Na figura 5.8 encontram-se representados os gráficos das médias do consumo alimentar médio diário por tercil de cada componente principal das várias vitaminas e sais minerais e das proteínas, disponíveis na base de dados original e selecionados neste estudo para sequente avaliação. Estes nutrientes encontram-se ajustados por energia.

Entre os antioxidantes aqui estudados, apontam-se a vitamina A (retinol), a vitamina C (ácido ascórbico), a vitamina E (alfa-tocoferol) e os carotenos. A vitamina A encontra-se em elevadas quantidades no fígado de vaca, porco ou frango, e também na cenoura, manteiga, pimento, couve, marisco e queijo. A vitamina C encontra-se em elevadas quantidades no pimento, tomate, alguns frutos (kiwi, morangos, citrinos, melão ou melancia), ice-tea e outros refrigerantes ou sumos de fruta/néctares embalados, hortícolas, batatas frita e fígado. Já a vitamina E encontra-se em grande abundância nos óleos vegetais, margarina, frutos secos, batatas fritas, salgados e ovos. Os carotenos são sintetizados pelas plantas, sendo encontrados na natureza sob duas formas principais: alfa-caroteno e beta-caroteno, apesar de existirem também outras formas. Estes consistem em pigmentos fotossintéticos, responsáveis pela cor de alguns vegetais (cenoura, pimento, couve, espinafres, brócolos, feijão verde) e frutos (tomate, melão, melancia, diospiro, pêssegos, ameixas), onde existem em abundância. Apresentam igualmente atividade antioxidante (Smiderle, 2013).

Quanto aos sais minerais estudados neste trabalho, tem-se o cálcio, ferro, magnésio e zinco. O cálcio pode ser encontrado em elevadas quantidades nos produtos lácteos (queijo, iogurte, leite), pizza (contém queijo), gelados de leite e sobremesas lácteas, alguns hortícolas (couve galega, grelos, espinafres,

nabiças), frutos secos, café, peixe gordo e marisco. O ferro, por sua vez, encontra-se em altos níveis nas carnes vermelhas (principalmente nas vísceras, como o fígado, tripas), incluindo bacon e hambúrguer, no marisco, frutos secos e, em menor quantidade, em alguns doces (chocolate, marmelada) e leguminosas. O magnésio e o zinco são também encontrados em elevada quantidade no chocolate e frutos secos, onde o magnésio é ainda encontrado em abundância no café e pão ou tostas integrais, e o zinco no queijo, carnes vermelhas (incluindo fiambre ou enchidos e hambúrguer), marisco, carne de perú ou coelho e pão ou tostas integrais.

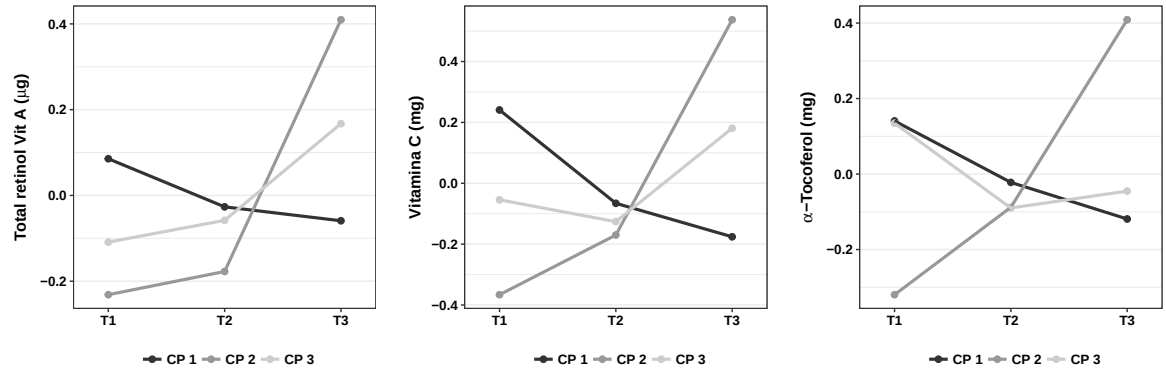
Finalmente, os produtos ricos em proteínas incluem os moluscos, crustáceos e peixe de conserva, as carnes brancas e vermelhas (incluindo fiambre ou enchidos, bacon ou toucinho e hambúrguer), peixe (gordo e magro, incluindo bacalhau), queijo, frutos secos e ovos.

O padrão "Saudável", tal como esperado, verifica um aumento acentuado para todos os nutrientes nos gráficos da figura 5.8 ao longo dos tercis. De facto, o consumo elevado de hortícolas (sopa), salada e fruta, característico deste padrão, justifica o elevado nível médio de antioxidantes ingeridos, bem como de sais minerais. O consumo elevado de iogurte e de peixe reflete-se igualmente no aumento da ingestão de proteínas e também de alguns sais minerais, como o cálcio.

Quanto ao padrão "Dieta tradicional portuguesa", uma vez que é um padrão que se caracteriza pelo elevado consumo de carnes vermelhas, bacalhau, pão e café é fácil perceber o aumento acentuado dos sais minerais ferro e magnésio e das proteínas ao longo dos tercis. Seria de esperar um aumento acentuado de zinco ao longo dos tercis, no entanto, do tercil de baixa adesão para o de moderada adesão observou-se uma diminuição dos valores de consumo médio desse nutriente, seguida de um aumento do tercil de moderada adesão para o de elevada adesão. Por último, foi verificado que os níveis médios dos antioxidantes, incluindo os carotenos, sofreram uma diminuição considerável ao longo dos tercis deste padrão, o que poderá ser explicado pela substituição das saladas, frutas e hortícolas, fontes principais desses nutrientes, pelo arroz, massa e batatas ou pão como acompanhamento.

O padrão "Petiscos"foi aquele que apresentou, no geral, níveis médios de nutrientes mais estáveis ao longo dos tercis, próximos da média zero. As alterações que aparentaram ser mais significativas ocorreram para o zinco, que apresentou um aumento considerável ao longo dos tercis, possivelmente justificado pelo elevado consumo de marisco neste padrão. O magnésio apresentou também um aumento do primeiro para o segundo tercil, seguido de uma diminuição para o terceiro tercil.

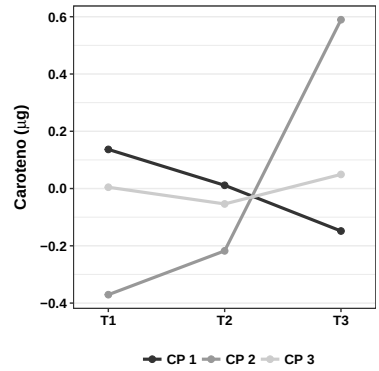
Por último, salienta-se que os indivíduos não são exclusivos a um padrão só, podendo ingerir diferentes alimentos de diferentes grupos alimentares. É, portanto, possível que os indivíduos pertencentes ao tercil de baixa adesão de determinado padrão em estudo possam estar a praticar maioritariamente outras condutas alimentares. Deste modo, o elevado consumo total de determinados nutrientes no primeiro tercil dum padrão poderá ser devido ao participante pertencer ao tercil de elevada adesão de outro padrão. Note-se, inclusivamente, que uma maior preferência ou adesão a um determinado padrão poderá resultar numa diminuição do consumo dos restantes padrões.



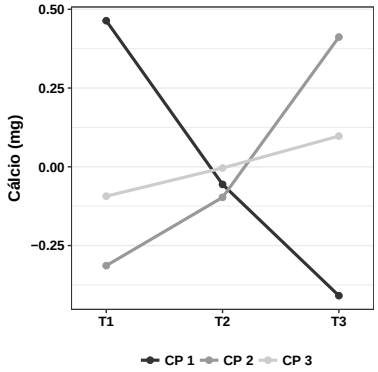
(a) Vitamina A (retinol)

(b) Vitamina C (ácido ascórbico)

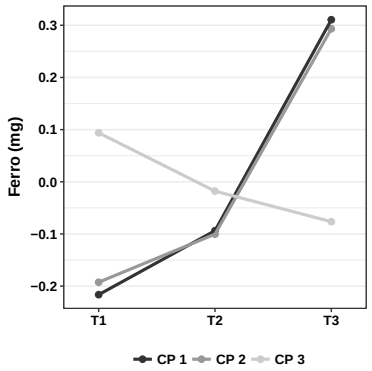
(c) Vitamina E (alfa-tocoferol)



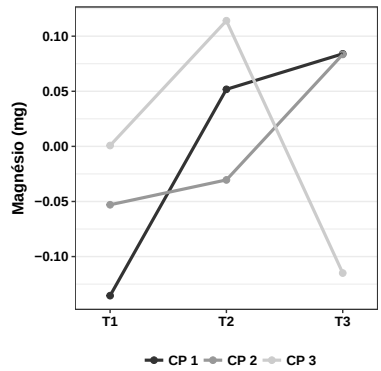
(d) Caroteno



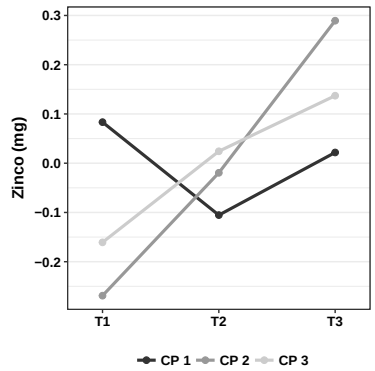
(e) Cálcio



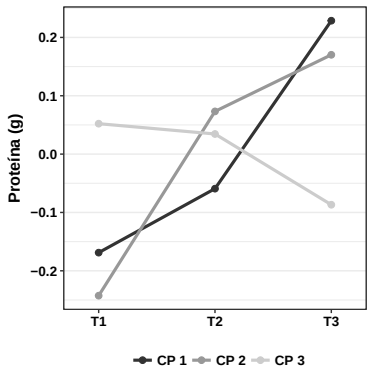
(f) Ferro



(g) Magnésio



(h) Zinco



(i) Proteínas

Figura 5.8: Gráficos das médias dos antioxidantes, sais minerais e proteínas (nutrientes ajustados para energia) no tercil (T) de cada componente principal (CP)

5.3 Regressão logística binária

Relevância clínica das variáveis de confundimento

Outro objetivo deste estudo foi avaliar o impacto dos padrões alimentares no desenvolvimento de DMI. Uma vez que outras variáveis poderão distorcer a verdadeira relação entre as variáveis de interesse (padrões alimentares) e o risco de DMI, procurou-se controlar dentro do possível o confundimento

existente no conjunto de dados, pela inclusão de variáveis clinicamente relevantes para a doença no modelo de regressão logística binária. Neste contexto, foi realizada uma revisão bibliográfica extensiva prévia ao processo de modelação, de modo a discernir quais as variáveis de confundimento a incluir no modelo. Nesta secção irão ser apresentados alguns dos principais resultados encontrados relativos à associação entre essas variáveis potencialmente relevantes e a DMI.

Tal como mencionado no capítulo introdutório, a idade avançada consiste no principal fator de risco de desenvolvimento de DMI, uma vez que o olho sofre diversas mudanças com o tempo que propiciam os processos fisiopatológicos inerentes à doença (Jager et al., 2008). De igual modo, o tabagismo, desde há muito conhecido pelo seu efeito acelerador do envelhecimento, é o fator de risco ambiental que demonstra maior consistência ao longo de vários estudos realizados, duplicando, aproximadamente, o risco de desenvolvimento de DMI e promovendo também a progressão da doença para a sua forma atrófica ou neovascular (Yu et al., 2016).

Alguns estudos epidemiológicos indicaram uma maior prevalência de DMI em mulheres comparativamente aos homens (Klein et al., 1992; Mitchell et al., 1995; Vingerling et al., 1995a), e um maior risco nas mulheres com menopausa precoce cirúrgica, após a remoção de um ou de ambos os ovários (Vingerling et al., 1995b). No *Coimbra Eye Study*, realizado para a população de Mira, uma zona costeira de Portugal, a prevalência da forma precoce de DMI foi maior para as mulheres, excetuando no grupo de maior idade ($\geq$85 anos). O mesmo não se verificou para as formas avançadas, que consoante o grupo etário apresentou uma prevalência maior ou menor para cada sexo (Cachulo et al., 2015).

Alguns autores observaram que indivíduos com valores de IMC fora do normal apresentaram risco aumentado da forma precoce e atrófica de DMI (Smith et al., 1998; Klein et al., 1998). No estudo transversal francês *Pathologies Oculaires Liées à l'Age* (POLA) foi igualmente encontrada uma relação entre as formas avançadas de DMI e anormalidades pigmentares em indivíduos com valores de IMC acima de 30 $\mathrm{kg/m^2}$ (Delcourt et al., 2001), à semelhança dos resultados obtidos por (Age-Related Eye Disease Study Research Group et al., 2000). Já no *Beaver Dam Eye Study* foi observada uma associação entre valores elevados de IMC e a forma precoce de DMI apenas nas mulheres, não se verificando esse efeito para os homens (Klein et al., 2001).

Em conjunto com o IMC, o perímetro abdominal e o rácio cintura-anca são também considerados medidas informativas de obesidade, uma vez que descrevem padrões de adiposidade abdominal, que poderão ter consequências metabólicas diferentes comparativamente a outros padrões de distribuição de gordura ou adiposidade global (como o IMC) (Seddon et al., 2003a). O perímetro abdominal tem sido associado a condições relacionadas com a obesidade, como a hipertensão e diabetes tipo II (Janssen et al., 2002), com outras doenças coronárias e cardiovasculares (Baik et al., 2000) e outros processos relacionados com a DMI, onde em vários estudos foi verificada uma associação positiva entre a obesidade abdominal com o desenvolvimento e progressão de DMI para as formas avançadas (Jager et al., 2008; Hawkins et al., 1999; Seddon et al., 2003a).

Os resultados do estudo *Hypertension, Cardiovascular Disease, and Age-Related Macular Degeneration* sugeriram que a forma neovascular de DMI se encontra associada a hipertensão sistémica moderada a severa, particularmente em doentes recebendo tratamento antihipertensivo (Hyman et al., 2000). Estes resultados foram consistentes com os de outros estudos (Kahn et al., 1977; Sperduto and Hiller, 1986; Klein and Klein, 1982; Delaney Jr and Oates, 1982), que reportaram uma associação da forma neovascular da doença com a hipertensão. Porém, noutros não foi encontrada qualquer

associação (Klein et al., 1997; Eye Disease Case-Control Study Group et al., 1992; Maltzman et al., 1979; Blumenkranz et al., 1986). Nenhum estudo identificou uma relação específica entre a hipertensão e a forma atrófica de DMI, levantando-se a hipótese de as formas avançadas da doença poderem ter uma patogénese distinta e de que a forma neovascular possa partilhar do processo sistémico subjacente à doença hipertensiva (Hyman et al., 2000).

A dislipidémia pode ser definida como um distúrbio nos níveis de lípidos e/ou lipoproteínas no sangue, caracterizada principalmente pela presença de valores elevados de colesterol. Esta consiste numa das principais doenças sistémicas e um dos principais fatores de risco cardiovascular. Devido ao seu impacto em vários órgãos do corpo, a dislipidémia tem sido direta e indiretamente associada a diversas doenças oculares, incluindo DMI, glaucoma, oclusão das veias retinianas e retinopatia (Wang et al., 2012).

O papel da diabetes no desenvolvimento e progressão da DMI tem sido também estudado por muitos investigadores. Alguns estudos reportaram uma associação positiva com a doença (Topouzis et al., 2009; Borger et al., 2003; Karesvuo et al., 2013; McGwin et al., 2003; Nitsch et al., 2008; Vaičaitienė et al., 2003), enquanto outros não demonstraram qualquer efeito (Fraser-Bell et al., 2008; Xu et al., 2009), ou até mesmo uma relação inversa (Clemons et al., 2006). Neste sentido, os autores do artigo *Diabetes Mellitus and Risk of Age-Related Macular Degeneration: A Systematic Review and Meta-Analysis* (Chen et al., 2014) realizaram uma meta-análise de modo a investigar essa associação, concluindo que a diabetes consiste, de facto, num fator de risco para DMI. A associação foi também de maior magnitude para as formas avançadas comparativamente à forma precoce.

Vários estudos sugeriram um risco aumentado das formas avançadas de DMI com o elevado consumo de bebidas alcoólicas (>40 g/dia) e uma fraca associação positiva entre o consumo alcoólico e a forma precoce da doença (Knudtson et al., 2007; Arnarsson et al., 2006; Buch, 2005; Fraser-Bell et al., 2006; Varma et al., 2004; Cho et al., 2000; Adams et al., 2012). Não existe evidência no sentido de que o consumo de baixos níveis de álcool desempenhe um efeito protetor de DMI, como reportado para as doenças cardiovasculares (Reynolds et al., 2003).

A prática de exercício físico tem sido reportada na literatura como tendo um efeito protetor para a DMI (Eye Disease Case-Control Study Group et al., 1992; Seddon et al., 2003a; Knudtson et al., 2006; Williams, 2009), embora os mecanismos adjacentes a esta associação não sejam ainda conhecidos. A redução do risco de doenças cardiovasculares, possivelmente preditoras de DMI, e a melhoria dos fatores de risco cardiovasculares (adiposidade, pressão arterial) que têm sido associados com o risco de DMI, embora de maneira não consistente, poderão estar na origem do efeito protetor que o exercício físico oferece (Williams, 2009).

Apesar de vários autores concluírem que o nível de escolaridade não se encontra associado à DMI (Klein et al., 1999; You et al., 2012), outros reportam uma associação com a forma precoce (Age-Related Eye Disease Study Research Group et al., 2000; Park et al., 2014). Vários estudos sugeriram que os participantes com maior nível educacional, são mais conscienciosos com a própria saúde e frequentemente com posições profissionais mais elevadas (Ederer et al., 1993), e que os hábitos tabágicos estão inversamente correlacionados com a realização educacional (Schottenfeld and Fraumeni Jr, 2006). Tal sugere também que a DMI possa ser afetada indireta ou diretamente por comportamentos e estilos de vida, assim como por fatores até ao momento não apurados.

Seleção de covariáveis

Neste projeto foi utilizado o método de análise de regressão logística binária para analisar a relação entre os padrões alimentares obtidos com o risco de DMI. A variável dependente foi `dmi`, que toma valor 0 caso o indivíduo não tenha sido diagnosticado com DMI e valor 1 caso contrário; e as variáveis independentes foram os *scores* dos indivíduos para os padrões alimentares obtidos por ACP (`pc1`, `pc2` e `pc3`), idade (`idade`), sexo (`sexo`), escolaridade(`esc_cat`), IMC (`imc`), perímetro abdominal (`perimetroabd`), hábitos tabágicos (`pacotesano` e `jafumador`), atividade física (`exercicio`), consumo energético total (`cia_energia_kcal`) e alcoólico (`alcool`), diabetes (`diabetes`), hipertensão (`hipertensao`) e dislipidémia (`dislipidemia`).

Numa fase preliminar, o conjunto das variáveis independentes foi analisado quanto à sua colinearidade, encontrando-se os valores GVIF calculados na tabela 5.5. É de salientar que algumas variáveis apresentam valores GVIF mais elevados, como por exemplo as variáveis `cia_energia_kcal`, `perimetroabd`, `imc`, `pc1`, `sexo` e `jafumador`, que à partida é expectável que estejam associadas (Mares et al., 2011). No entanto, todos os valores GVIF foram inferiores a 3.16 ($10^{\frac{1}{2}}$) e, no caso da variável `esc_cat`, a 1.72 ($10^{\frac{1}{2\times 2}}$), sendo de esperar que a multicolinearidade não afete os resultados do modelo.

Tabela 5.5: Valores $\text{GVIF}^{1/(2*\text{g.l.})}$ das covariáveis candidatas ao modelo estimado (graus de liberdade, g.l.)

Variável	**g.l.**	$\textbf{GVIF}^{\mathbf{1/(2*g.l.)}}$
`pc1`	1	1.542
`pc2`	1	1.284
`pc3`	1	1.276
`idade`	1	1.107
`imc`	1	1.613
`cia_energia_kcal`	1	1.895
`sexo`	1	1.476
`perimetroabd`	1	1.725
`exercicio`	1	1.028
`diabetes`	1	1.056
`hipertensao`	1	1.075
`dislipidemia`	1	1.036
`alcool`	1	1.314
`jafumador`	1	1.459
`pacotesano`	1	1.313
`esc_cat`	2	1.031

Antes da construção do modelo de regressão logística binária é importante discutir um tipo especial de variável que ocorre frequentemente na prática. Note-se que a variável `pacotesano` tem valor 0 para todos os não fumadores, e que essa ocorrência sucede com uma frequência muito maior do que a esperada para uma distribuição completamente contínua. Adicionalmente, os valores diferentes de 0 exibem assimetria à direita (ver figura 5.9). Em 1994, Robertson, Boyle, Hsieh, Macfarlane e Maisonneuve demonstraram que a maneira correta de modelar uma variável com estas características é incluir dois termos: uma variável dicotómica com valor 0 para não fumadores e 1 para fumadores (aqui denominada de `jafumador`) e a variável original, que contém os valores observados (`pacotesano`) (Robertson et al., 1994). Deste modo, o *logit* para um modelo univariado é

$$logit(p_i) = \beta_0 + \beta_1 \texttt{jafumador} + \beta_2 \texttt{pacotesano}, \tag{5.1}$$

onde `jafumador`=0 se `pacotesano`=0 e `jafumador`=1 se `pacotesano`>0. É também importante

acrescentar que durante o processo de modelação é ainda necessário verificar a escala do *logit* para os valores positivos da covariável.

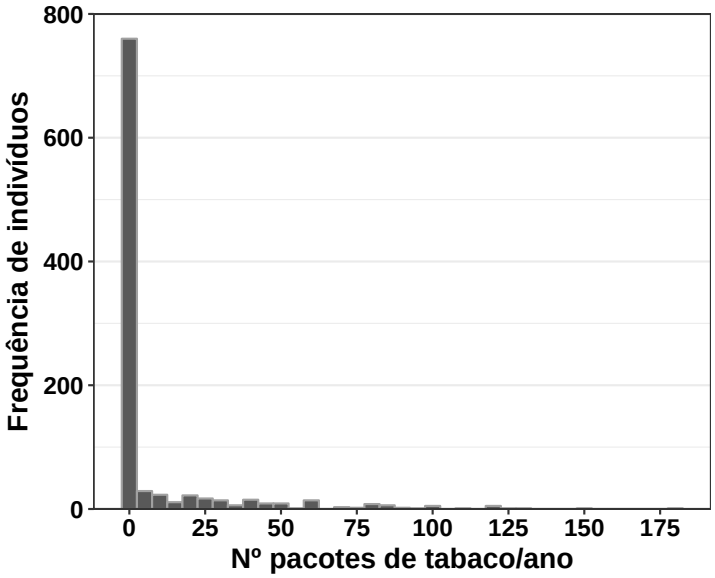


Figura 5.9: Histograma da distribuição dos indivíduos por nº pacotes de tabaco consumidos anualmente

Previamente ao processo de modelação foi realizada uma análise univariada para cada variável, onde se ajustaram modelos univariados para cada uma das covariáveis. Na tabela 5.6 encontram-se descritas as estimativas dos parâmetros e os respetivos erros padrão resultantes do ajustamento dos modelos de regressão logística simples. Encontram-se igualmente descritos os valores-p do teste de Wald (valor-p Wald) e do teste da razão de verosimilhanças entre o modelo logístico univariado e o modelo nulo (valor-p RV). Adicionalmente, para as variáveis categóricas, foram construídas e analisadas tabelas de contingência do valor observado para a variável dependente (`dmi`=0,1) contra os k níveis da variável independente em questão. Estas tabelas não estão aqui representadas, uma vez que as frequências destas variáveis já se encontram descritas na tabela 5.2.

Para as variáveis contínuas a análise foi suplementada com a construção de gráficos de dispersão $smooth$ (para uma discussão mais completa de métodos gráficos de *smoothing*, ver (Cleveland and Loader, 1996)), de modo a determinar não só a potencial importância da variável, mas também a possível presença e impacto de observações extremas, e a escala apropriada da variável. Estes gráficos foram utilizados como suporte no processo de modelação, onde na figura 5.10 se encontra um exemplo de um gráfico *lowess* para a variável `perimetroabd`.

Tabela 5.6: Estimativas dos parâmetros, erros padrão e valores-p dos modelos de regressão logística univariados

Covariável	**Estimativa coeficiente**	**Erro padrão**	**valor-*p* Wald**	**valor-*p*^{a} RV**
`pc2`	-0.1509	0.0482	0.0017***	0.001***
`cia_energia_kcal`	-0.0002	0.0001	0.0950*	0.094*
`exercicio`	-0.1917	0.1435	0.1817	0.181
`perimetroabd`	-0.0062	0.0048	0.1979	0.196
`hipertensao`	0.1671	0.1401	0.2332	0.232
`diabetes`	-0.1707	0.1478	0.2481	0.247
`pc3`	-0.0539	0.0486	0.2668	0.264
`pc1`	0.0488	0.0440	0.2670	0.267
`idade`	0.0077	0.0083	0.3551	0.355
`imc`	-0.0132	0.0149	0.3744	0.374
`esc_cat(6,9]`	-0.1375	0.1714	0.4226	0.399
`esc_cat(9,16]`	-0.3026	0.2237	0.1761	
`sexo`	-0.0639	0.1303	0.6240	0.624
`dislipidemia`	-0.0549	0.1302	0.6733	0.673
`alcool`	0.0013	0.0046	0.7729	0.773
`jafumador`	-0.1162	0.1910	0.5429	0.829
`pacotesano`	0.0016	0.0038	0.6701	

Significância estatística aos níveis α=0.1 (*) e α=0.001 (***)

^{a} As variáveis encontram-se ordenadas por ordem decrescente de significância estatística.

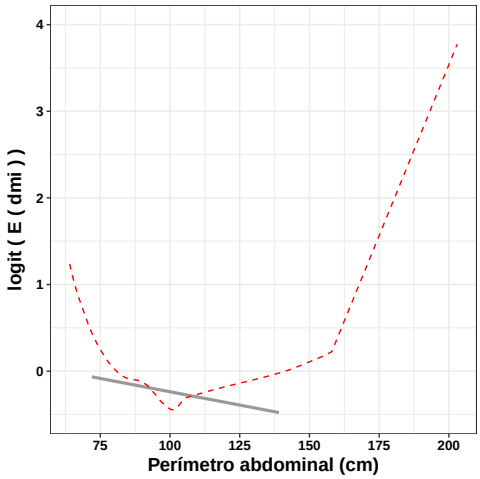


Figura 5.10: Gráfico do *logit* para a variável `perimetroabd` (linha cinzenta-modelo logístico univariado, linha tracejada a vermelho-curva *lowess*)

A variável `pc2` foi a única que aparentou ser importante para a explicação da variável resposta ao nível de significância de $\alpha = 0.05$. A variável `cia_energia_kcal` apresentou também significância estatística, porém ao nível de significância $\alpha = 0.1$. É de salientar que nenhuma das variáveis categóricas do estudo apresentou células vazias na respetiva tabela de contingência, não existindo um condicionamento de uma estimativa pontual dos OR a tender ou para zero ou para infinito. Note-se também que todos os valores esperados em cada célula foram superiores a 5 (dados aqui não demonstrados). Pode verificar-se que os valores-p dos testes de Wald e RV são semelhantes para as variáveis contínuas ou categóricas de dois níveis, uma vez que o tamanho da amostra é grande e estes testes são assintoticamente equivalentes. No entanto, para a variável `esc_cat`, que tem mais de dois níveis, reporta-se apenas um valor-p global no teste RV para a variável e vários valores-p Wald para

cada nível da variável categórica (exceptuando para o de referência). Tal acontece porque o teste RV é realizado com base nos modelos com a variável e nulo, testando o seu efeito geral, enquanto que o teste de Wald é realizado para cada nível da variável categórica, permitindo testar se a diferença entre o coeficiente do nível de referência e os coeficientes dos outros níveis é igual a zero. Pela análise da tabela, podemos assim verificar que os níveis `(6,9]` e `(9,16]` não diferem significativamente do nível de referência `(1,6]`, e que, no geral, a variável não é significativa no modelo (valor-p RV$= 0.399$). Acrescenta-se também que, alterando os níveis de referência, não se reportaram quaisquer diferenças significativas entre esses e os restantes níveis (resultados não demonstrados). Para as variáveis `jafumador` e `pacotesano` foram realizados testes de Wald individualmente, e teste RV entre o modelo com as duas variáveis e o modelo nulo.

Para todas as variáveis contínuas foi verificado o pressuposto de linearidade do *logit* através de gráficos *lowess*, no entanto, dado o seu grande número, apenas se encontra aqui presente o gráfico da figura 5.10 para a variável `perimetroabd`. Através da sua análise, podemos verificar que o pressuposto de linearidade do *logit* para `perimetroabd`, assumido no modelo logístico univariado (linha cinzenta), não é cumprido, uma vez que a curva *lowess* (linha tracejada a vermelho) parece ter uma forma quadrática. Esta informação foi útil, portanto, na etapa da determinação da verdadeira forma funcional da variável contínua do processo de modelação.

Independentemente do facto de não terem apresentado significância na análise univariada, foram selecionadas todas as variáveis com potencial relação com a DMI (descritas na secção 5.3) para a construção de um modelo.

• ***Modelo I***

Foi obtido um primeiro modelo de regressão logística binária (*Modelo I*), pela aplicação da metodologia descrita na secção 5.3. Este modelo não consistiu no modelo final, pois apresentou observações influentes que foram subsequentemente eliminadas e foi gerado um novo modelo (*Modelo II*). Por esta razão não se achou relevante aqui descrever detalhadamente o seu processo de construção. No entanto, serão apresentadas medidas sumárias da qualidade de ajustamento do modelo e resultados da análise de resíduos efetuada.

O *Modelo I* é dado pela seguinte equação:

$$\begin{aligned} logit[E(\texttt{dmi}_i)] = {} & 0.0762 + 0.1323\,\texttt{pc1}_i - 0.0972\,\texttt{pc2}_i + 0.0267\,\texttt{pc3}_i + 0.0020\,\texttt{idade}_i \\ & - 0.0003\,\texttt{cia\_energia\_kcal}_i + 0.2170\,\texttt{hipertensao}_i - 0.1572\,\texttt{exercicio}_i \\ & - 0.1212\,\texttt{sexo}_i - 3.4088\,\texttt{perimetroabd}_i + 4.7251\,\texttt{perimetroabd}_i^2 \\ & - 0.1344\,\texttt{dislipidemia}_i + 0.0431\,\texttt{alcool}_i - 0.1062\,S_{A,3,1,i} + 0.3376\,\texttt{pacotesano}_i \\ & - 48.4544\,S_{P,5,1,i} + 55.7185\,S_{P,5,2,i} - 8.1480\,S_{P,5,3,i} - 0.7445\,\texttt{jafumador}_i, \end{aligned} \tag{5.2}$$

onde a componente RCS associada à variável `alcool` se define por

$$S_{A,3,1} = (\texttt{alcool}_i - 0)_+^3 - \frac{26.5-0}{26.5-4}(\texttt{alcool}_i - 4)_+^3 + \frac{4-0}{26.5-4}(\texttt{alcool}_i - 26.5)_+^3 \tag{5.3}$$

e as componentes RCS associadas à variável `pacotesano` se definem por

$$
\begin{aligned}
S_{P,5,1} &= (\texttt{pacotesano}_i - 0.05)_+^3 - \tfrac{100-0.05}{100-40}(\texttt{pacotesano}_i - 40)_+^3 + \tfrac{40-0.05}{100-40}(\texttt{pacotesano}_i - 100)_+^3 \\
S_{P,5,2} &= (\texttt{pacotesano}_i - 1.98)_+^3 - \tfrac{100-1.98}{100-40}(\texttt{pacotesano}_i - 40)_+^3 + \tfrac{40-1.98}{100-40}(\texttt{pacotesano}_i - 100)_+^3 \\
S_{P,5,3} &= (\texttt{pacotesano} - 16.6)_+^3 - \tfrac{100-16.6}{100-40}(\texttt{pacotesano} - 40)_+^3 + \tfrac{40-16.6}{100-40}(\texttt{pacotesano} - 100)_+^3 .
\end{aligned}
\tag{5.4}
$$

Note-se que a variável `perimetroabd` se encontra inserida do modelo sob a forma de polinómio de 2º grau, segundo o mencionado na subsecção anterior.

Para avaliar a qualidade de ajustamento do modelo foi aplicado o teste da soma não ponderada de quadrados, onde se obteve um valor-p de 0.21. Como tal, não existiu evidência para rejeitar a hipótese nula de que as verdadeiras probabilidades são aquelas especificadas pelo modelo.

De seguida, procedeu-se à análise de resíduos para verificar se o bom ajustamento era suportado ao longo do conjunto total de padrões das covariáveis. Pela análise do gráfico da esquerda da figura 5.11, verificou-se que nenhum resíduo de Pearson estandardizado teve um valor absoluto superior 2, não sendo identificados, à partida, potenciais *outliers*. Observou-se também que a curva *lowess* incorporada no gráfico resultou aproximadamente numa linha horizontal com interceção em 0, exceto nos valores mais extremos de *logit*, o que sugere que o modelo está correto e não existem *outliers* significativos.

Na tabela 5.7 estão presentes os valores dos resíduos de Pearson estandardizados (prs_i), de alavacagem (h_{ii}) e distância de Cook ($\Delta\beta_i$) que verificaram maior magnitude no conjunto de dados. O gráfico da direita da figura 5.11 ilustra os resíduos estandardizados em função da alavancagem, bem como a distância de Cook para cada um dos valores observados da variável resposta, destacando os valores com maior efeito sobre as estimativas dos parâmetros.

Pode observar-se que o participante com ID 2443 tem os maiores valores de alavancagem e de distância de Cook, seguido pelo participante com ID 1365 com o segundo maior valor de alavancagem e de distância de Cook. O participante com ID 1268 detém o quarto maior valor de alavancagem, apresentando um valor da distância de Cook semelhante aos outros dois participantes, sendo assim, de influência semelhante. Já o participante com ID 967, apesar de ter um elevado valor de alavancagem, apresenta uma menor distância de Cook comparativamente às outras observações exercendo, deste modo, menor influência no modelo.

Quando os indivíduos com ID 2443, 1268 e 967 foram retirados individualmente, verificaram-se alterações acima de 10% nos coeficientes estimados. Para o indivíduo de ID 1365 não se registaram alterações significativas nos coeficientes. Foram subsequentemente comparados os coeficientes do modelo reajustado após a eliminação simultânea dos participantes de ID 2443, 1268 e 967 com os coeficientes do modelo reajustado retirando as quatro observações. De facto, a eliminação adicional do participante de ID 1365 resultou numa alteração de 20% no coeficiente associado à variável `pc3`.

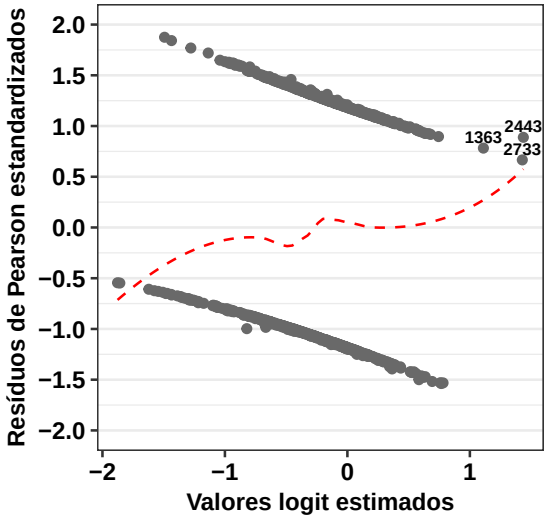
Quando estudadas estas observações mais detalhadamente, verificamos que o participante de ID 2443 é uma mulher com DMI que apresentou os valores máximos de idade (94 anos) e de perímetro abdominal (203 cm). Apresentou também valores de IMC, de consumo energético médio e *scores* para o padrão "Saudável"e para o padrão "Petiscos"acima do terceiro quartil. Seria de esperar que esta participante praticasse um estilo de vida sedentário e que apresentasse problemas associados ao excesso de peso, no entanto, esta reportou ser praticante regular de exercício físico e não padecer de qualquer uma das comorbilidades avaliadas no estudo. Tal demonstra que, efetivamente, esta observação

é incomum, decidindo-se pela sua eliminação.

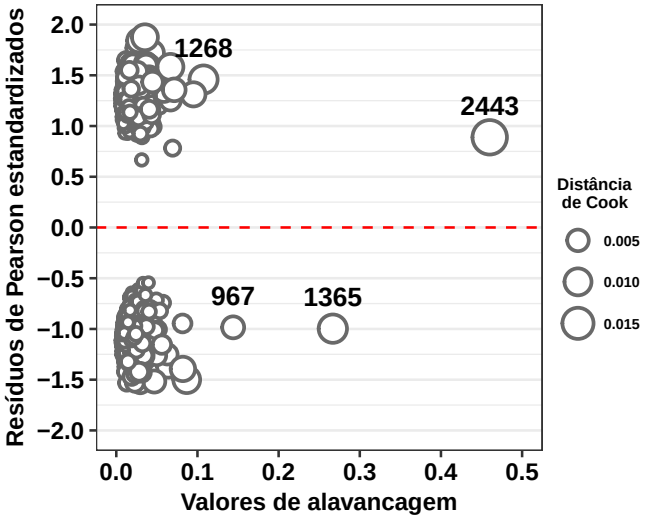
Por sua vez, o indivíduo de ID 1268, do sexo feminino e com DMI, apresentou o valor máximo de *score* (11.92) para o padrão "Petiscos", valores baixos para os outros padrões e um consumo energético acima do terceiro quartil. Os indivíduos de ID 1365 e 967 apresentaram os dois maiores valores de pacotes de tabaco consumidos por ano. O primeiro sem DMI e com valores de perímetro abdominal, consumo energético médio, consumo alcoólico e *scores* para os padrões "Dieta Tradicional Portuguesa"e "Petiscos"acima do terceiro quartil, enquanto que o segundo apresentou valores de idade e de *score* para o padrão "Saudável"acima do terceiro quartil e "score"para o padrão "Dieta Tradicional Portuguesa"abaixo do primeiro quartil.

Tabela 5.7: Valores dos resíduos de Pearson estandardizados (prs_i), de alavacagem (h_{ii}) e distância de Cook ($\Delta\beta_i$)

ID	rps_i	h_{ii}	$\Delta\beta_i$
358	1.7729738	0.02588096	0.005154355
967	-0.9567196	0.14399520	0.005311651
1268	1.4456247	0.10750653	0.011254697
1365	-0.9429223	0.26679909	0.011488752
1726	1.8492267	0.02783646	0.006518487
2066	1.8852573	0.03527894	0.008879212
2443	0.7931673	0.46008153	0.019746645



(a) Gráfico dos resíduos de Pearson estandardizados em função dos valores *logit* estimados. A linha tracejada vermelha representa a curva *lowess smooth* dos resíduos contra as probabilidades logísticas estimadas.



(b) Gráfico de influência proporcional dos resíduos de Pearson estandardizados em função dos valores de alavancagem. O tamanho dos círculos é proporcional à distância de Cook apresentada para as respetivas observações.

Figura 5.11: Gráficos utilizados no diagnóstico do *Modelo I*

• ***Modelo II***

O processo de construção do modelo final (*Modelo II*) partiu dum modelo inicial onde todas as variáveis contínuas entram no modelo com uma componente *restricted cubic splines* de 5 nós (ver secção 5.3). Numa fase inicial foi testada a não linearidade das variáveis de interesse do estudo, isto é,

referentes aos padrões alimentares (`pc1`, `pc2` e `pc3`), por ordem crescente de significância no modelo inicial. Na tabela 5.8 encontram-se os valores de AIC para cada modelo com as variáveis de interesse na sua forma linear e nas diferentes formas não lineares; e os valores-p referentes aos testes da razão de verosimilhança entre o modelo linear e os modelos não lineares.

Tabela 5.8: Valores de AIC e testes de não linearidade para modelo com diferentes formas das variáveis de interesse

	`pc3`		`pc1`		`pc2`	
	AIC	**valor-p RV**	**AIC**	**valor-p RV**	**AIC**	**valor-p RV**
Linear	1348.287	-	1343.099	-	1340.758	-
Restricted cubic splines (N^o. nós)						
3	1349.138	0.284	1345.099	0.989	1342.306	0.501
4	1350.975	0.519	1346.988	0.946	1341.393	0.186
5	1352.965	0.724	1348.287	0.847	1343.099	0.301
6	1353.855	0.657	1344.393	0.152	1343.696	0.281
7	1355.276	0.698	1344.734	0.137	1343.367	0.193
8	-	-	1345.919	0.164	-	-
9	-	-	-	-	-	-
Polinomial						
2^o grau	1349.799	0.485	1344.838	0.610	1342.627	0.717
3^{o}grau	1351.731	0.757	1346.822	0.871	1341.639	0.210

A variável `pc3` foi a primeira a ser testada. Pode-se verificar pela análise da tabela que, entre os modelos com *splines*, o melhor foi aquele com 3 nós, pois apresentou o menor valor de AIC. Este modelo verificou também ser melhor que os polinomiais, no entanto, o valor de AIC foi maior do que o do modelo linear. Adicionalmente, ao serem analisados os valores-p dos testes da razão de verosimilhanças entre os vários modelos não lineares e o modelo linear confima-se que, de facto, os modelos não são significativamente diferentes (valor-$p > 0.05$), optando-se pelo modelo de menor complexidade, ou seja, o modelo linear.

Procedeu-se ao reajustamento do modelo inicial, agora com a variável na forma linear, repetindo-se todo o processo acima descrito para a segunda variável de interesse com menor significância no modelo, a variável `pc1`. Mais uma vez, para esta variável não se verificou evidência estatística de não linearidade, o mesmo acontecendo para a última variável de interesse testada, a variável `pc2`.

A tabela 7.14 do Apêndice C apresenta os vários passos da eliminação *backward* das potenciais variáveis de confundimento, onde se encontram os valores-p dos testes de razão de verosimilhança entre o modelo ajustado com e sem a variável em questão e as percentagens das diferenças dos coeficientes das variáveis de interesse do modelo sem a variável e do modelo completo.

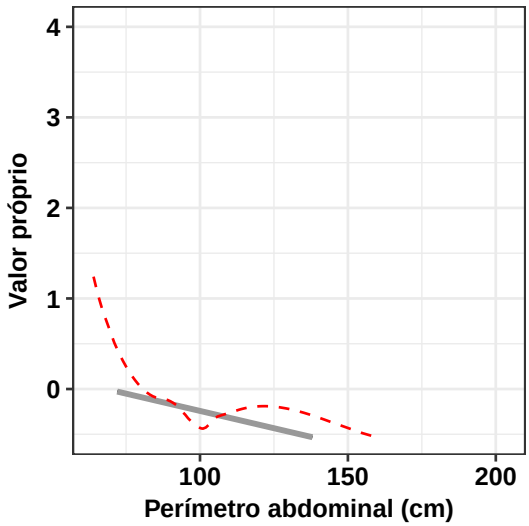
Como se pode observar, apenas as variáveis `esc_cat` e `diabetes` foram eliminadas do modelo, não só porque o valor-p do teste de razão de verosimilhanças entre o modelo com e sem a variável foi maior do que 0.20, como também não se verificou uma alteração significativa ($> 10\%$) dos coeficientes das variáveis de interesse comparativamente aos do modelo corrente.

O passo sequente consistiu em testar a não linearidade das variáveis de confundimento contínuas do modelo, por um processo análogo ao anteriormente descrito para as variáveis de interesse. Uma vez que

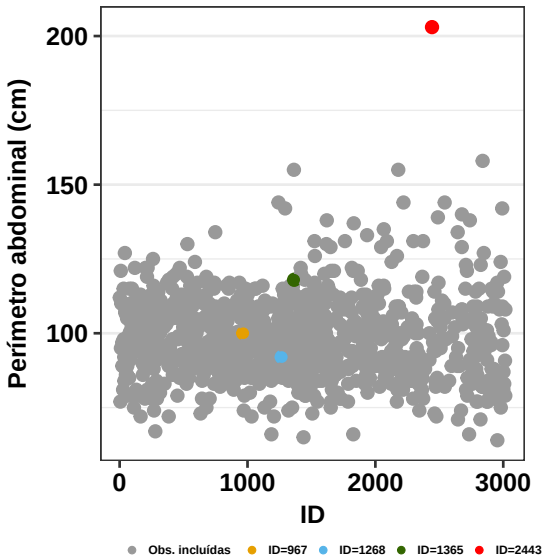
existe um elevado número de variáveis a testar, esse processo não será aqui descrito exaustivamente. Na tabela 7.15 do Apêndice C encontram-se os valores de AIC para cada modelo construído com diferentes formas da variável contínua a ser testada, bem como os valores-p dos testes de razão de verosimilhanças entre o modelo com a variável na sua forma linear e com a variável em diferentes formas não lineares. Note-se que as variáveis são sempre testadas por ordem crescente de significância estatística no modelo reajustado.

Pela análise da tabela verifica-se que as variáveis `imc`, `idade`, `cia_energia_kcal` e `perimetroabd` apresentaram valor-$p > 0.05$ dos testes da razão de verosimilhanças entre os vários modelos não lineares e o modelo linear, optando-se reajustar o modelo com as variáveis na sua forma linear. O modelo com a variável `alcool` na forma RCS com 3 nós apresentou menor valor de AIC relativamente a todas as outras formas e também um valor-$p \leq 0.05$ no teste RV comparando com a forma linear da variável, pelo que se reajustou o modelo com a variável `alcool` e as suas componentes RCS de 3 nós associadas. O mesmo sucedeu para a variável `pacotesano`, porém para a a forma RCS com 5 nós, reajustando-se o modelo com essa forma da variável.

É curioso verificar que a variável `perimetroabd` se encontra agora, no *Modelo II*, sob a forma linear. De facto, após a exclusão dos indivíduos de ID 2443, 1268, 967 e 1365, foram novamente construídos gráficos *lowess* para cada modelo univariado. O gráfico assim obtido para `perimetroabd` encontra-se na figura 5.12 à esquerda. Pela sua observação, constatou-se que a forma funcional desta variável aproximou-se mais da forma linear por causa da remoção da observação de ID=2443, que por apresentar um valor excessivamente alto comparativamente às restantes observações, estava a influenciar o gráfico *lowess* da figura 5.10, conferindo-lhe uma forma quadrática (ver gráfico de dispersão à direita da figura 5.12).



(a) Gráfico dos valores *logit* estimados para a variável `perimetroabd` e sem observações de ID 2443, 1268, 967 e 1365 (linha cinzenta: modelo logístico univariado; linha tracejada a vermelho: curva *lowess*)



(b) Gráfico de dispersão dos valores de perímetro abdominal por cada observação identificada por ID

Figura 5.12: Estudo da linearidade do *logit* para a variável `perimetroabd`

O último passo no processo de modelação foi a determinação da existência de interações entre as

covariáveis do modelo (ver tabela 7.16, Apêndice C). Apesar de algumas interações terem apresentado significância estatística ao nível $\alpha = 0.05$, como as interações entre sexo e perímetro abdominal (valor-$p = 0.039$), entre idade e exercício (valor-$p = 0.024$) e entre o segundo padrão alimentar e o número de pacotes de tabaco consumidos anualmente (valor-$p = 0.028$), optou-se por não se incluir nenhuma no modelo final. Tal decisão fundamentou-se nomeadamente no facto de estas interações não se encontrarem reportadas na literatura e não fazerem sentido de um ponto de vista biológico ou clínico.

O modelo final estimado (*Modelo II*) foi então:

$$\begin{aligned} logit[E(\texttt{dmi}_i)] = {} & 0.9408 + 0.1345\,\texttt{pc1}_i - 0.1121\,\texttt{pc2}_i - 0.0019\,\texttt{pc3}_i + 0.0018\,\texttt{idade}_i \\ & - 0.0002\,\texttt{cia\_energia\_kcal}_i - 0.1397\,\texttt{exercicio}_i - 0.1047\,\texttt{sexo}_i \\ & - 0.0141\,\texttt{perimetroabd}_i + 0.0147\,\texttt{imc}_i - 0.1292\,\texttt{dislipidemia}_i \\ & + 0.0441\,\texttt{alcool}_i - 0.1091\,S_{(A,3,1,i)} + 0.3387\,\texttt{pacotesano}_i - 50.5555\,S_{(P,5,1,i)} \\ & + 58.2002\,S_{(P,5,2,i)} - 8.3037\,S_{(P,5,3,i)} - 0.7424\,\texttt{jafumador}_i + 0.2152\,\texttt{hipertensao}_i, \end{aligned} \tag{5.5}$$

onde as componentes RCS associadas às variáveis `alcool` e `pacotesano` já se encontram previamente definidas (ver *Modelo I*).

Na tabela 5.9 estão apresentadas as estimativas dos coeficientes do *Modelo II* e respetivos erros padrão, assim como o valor-p referente ao teste de Wald, a estimativa do OR e a correspondente estimativa do intervalo de 95% de confiança.

Tabela 5.9: Estimativas dos parâmetros, erros padrão, valores-p do teste de Wald, razão de chances (OR) e respetivos intervalos de 95% de confiança do *Modelo II*

Covariável	**Estimativa coeficiente**	**Erro padrão**	**valor-*p* Wald**	$\boldsymbol{OR}$	**IC 95% $\boldsymbol{OR}$ (Lim. Inferior, Superior)**
`(Intercept)`	0.9408	1.0069	0.3501	2.562	(0.356, 18.436)
`pc1`	0.1345	0.0703	0.0559*	1.144	(0.997, 1.313)
`pc2`	-0.1121	0.0635	0.0773*	0.894	(0.789, 1.012)
`pc3`	-0.0019	0.0666	0.9775	0.998	(0.876, 1.137)
`idade`	0.0018	0.0093	0.8508	1.002	(0.984, 1.020)
`cia_energia_kcal`	-0.0002	0.0003	0.3500	1.000	(0.999, 1.000)
`exercicio`	-0.1397	0.1519	0.3580	0.870	(0.646, 1.171)
`sexo`	-0.1047	0.1966	0.5945	0.901	(0.613, 1.324)
`perimetroabd`	-0.0141	0.0088	0.1102	0.986	(0.969, 1.003)
`imc`	0.0147	0.0248	0.5524	1.015	(0.967, 1.065)
`dislipidemia`	-0.1292	0.1375	0.3476	0.879	(0.671, 1.151)
`alcool`	0.0441	0.0215	0.0405**	-	-
$S_{(A,3,1)}$	-0.1091	0.0529	0.0390**	-	-
`pacotesano`	0.3387	0.1163	0.0036***	-	-
$S_{(P,5,1)}$	-50.5555	17.9483	0.0049***	-	-
$S_{(P,5,2)}$	58.2002	20.7510	0.0050***	-	-
$S_{(P,5,3)}$	-8.3037	3.1532	0.0085***	-	-
`jafumador`	-0.7424	0.3781	0.0496**	0.476	(0.227, 0.999)
`hipertensao`	0.2152	0.1531	0.1599	1.240	(0.919, 1.674)

Significância estatística ao níveis $\alpha = 0.1$ (*), $\alpha = 0.05$ (**) e $\alpha = 0.01$ (***)

Antes de se proceder à interpretação do modelo final obtido, é importante relembrar que um dos objetivos principais deste trabalho foi quantificar e compreender a potencial relação entre os padrões

alimentares e o risco de DMI, e não o impacto das potenciais variáveis de confundimento selecionadas na prevalência da doença. Deste modo, focou-se exclusivamente na interpretação das variáveis `pc1`, `pc2` e `pc3`, que foi feita com base no OR, tendo em atenção que os indivíduos apenas diferem na característica a ser analisada, partilhando os mesmos valores para as restantes covariáveis. Tal está patente na tabela 5.9, onde as variáveis de confundimento se encontram a cinzento.

As variáveis de interesse `pc1` e `pc2` não contribuem significativamente para o modelo ao nível $\alpha = 0.05$. Porém, contribuem significativamente ao nível $\alpha = 0.1$, apresentando valores-p no teste de Wald de 0.056 e 0.077, respetivamente. De facto, para o *score* do padrão "Saudável"(`pc2`) observou-se uma associação protetora em respeito à DMI, onde o OR obtido de 0.894 pode ser interpretado como uma diminuição esperada de 10.6% na chance de um indivíduo apresentar DMI com o aumento de uma unidade do *score*. Já o *score* para o padrão "Dieta Tradicional Portuguesa"(`pc1`) teve um impacto positivo na doença, onde o OR obtido de 1.144 indica um aumento esperado de 14.4% da chance de ter a doença por aumento de unidade do *score*. Por fim, o padrão "Petiscos"(`pc3`) não revelou qualquer associação com o risco de DMI, uma vez que não só se verificou um coeficiente estimado associado a `pc3` próximo de zero (-0.0019), como também este demonstrou não ser estatisticamente significativo, com um valor-p de 0.978 no teste de Wald e um OR próximo de 1 (0.998).

Quanto às variáveis de confundimento, as variáveis `jafumador`, e as variáveis `alcool` e `pacotesano` e componentes RCS associadas demonstraram evidência de significância estatística aos níveis de significância usuais. Note-se que seria de esperar um coeficiente positivo para a variável `jafumador`, dado o tabagismo ser o fator de risco para DMI mais consistente ao longo dos estudos, como já mencionado. A interpretação do valor OR de 0.476 poderia levar-nos a afirmar que a propensão para o indivíduo ser doente entre os fumadores ou ex-fumadores tem uma diminuição esperada de 52.4% relativamente aos indivíduos não fumadores. No entanto, saliente-se novamente que a variável `pacotesano` e `jafumador` têm de ser analisadas conjuntamente. Pressuponha-se que um determinado indivíduo fuma em média 1 cigarro por semana, o que resulta em 2.6 pacotes de tabaco consumidos por ano. Comparativamente a um outro indivíduo não fumador, esse fumador teria, em média, uma chance de ter a doença superior em 15.1% ($exp(2.607 \times 0.3387 - 0.7424 \times 1) = 1.151$), mantendo as outras variáveis constantes. Resta-se também acrescentar que existem poucos fumadores e ex-fumadores na amostra do estudo (26.16%), onde entre os indivíduos sem DMI esses representam 14.89% dos casos e entre os indivíduos com DMI 11.27%. Além disso, quando analisado o número de pacotes de tabaco consumidos por ano entre esses participantes, verificou-se que em média apenas 8.23 e 8.39 pacotes foram consumidos entre os indivíduos com e sem DMI, respetivamente, o que se traduz numa média inferior a um cigarro fumado por dia. Deste modo, a maioria dos participantes fumadores ou ex-fumadores eram fumadores esporádicos, sendo que este grupo poderá não ser representativo da população de interesse e não reunir as mesmas condições tabágicas sob as quais outros estudos demonstraram uma associação positiva entre o risco da doença e o tabagismo.

No gráfico da esquerda da figura 7.7 do Apêndice C encontram-se representados os valores *logit* estimados segundo a variável `alcool` e componentes RCS associadas (mantendo as restantes variáveis do *Modelo II* constantes). Pela sua análise, verifica-se um aumento do *logit* até ao consumo médio diário de álcool de 15 ml e, a partir desse valor, uma diminuição aproximadamente linear do *logit*. Note-se que o valor 15 ml reporta a 1.3 copos de vinho, 1.23 latas ou garrafas de cerveja e 0.85 cálices de bebidas brancas ou espirituosas por dia, ou seja, aproximadamente 1 bebida alcoólica diária. Vários estudos sugeriram que não só a quantidade, mas também o tipo de álcool poderá ser importante: o consumo de cerveja encontrou-se associado com o aumento do risco de DMI (Ritter et al., 1995;

Moss et al., 1998), enquanto que o consumo de vinho, com a diminuição do risco da doença (Moss et al., 1998; Arnarsson et al., 2006; Obisesan et al., 1998). Deste modo, foi construído um boxplot (gráfico da direita) das distribuições do consumo alcoólico médio diário segundo as diferentes bebidas alcoólicas. Note-se que o teor alcoólico de cada bebida foi calculado segundo a Tabela da Composição de Alimentos (INSA, 2006), considerando-se 1 g de álcool equivalente a 1 ml de álcool para os cálculos. Pela respetiva análise verifica-se uma grande dispersão dos valores de consumo alcoólico referentes ao vinho comparativamente às outras bebidas alcoólicas, para as quais a maioria dos indivíduos não reportou qualquer consumo, e poucos referiram consumir até 1 ou 2 bebidas (no caso da cerveja) por dia. Tal sugere que o efeito protetor (diminuição do *logit* com o aumento do consumo de álcool) poderá ser exercido em grande parte pelo consumo de vinho, à semelhança dos resultados obtidos nos estudos acima referidos. Saliente-se também que um indivíduo poderá não só consumir apenas um tipo de bebida alcoólica, não se podendo concluir que o efeito protetor seja devido unicamente ao vinho, ou a diferentes combinações de várias bebidas, dado que não se pode discernir o tipo de álcool no gráfico dos valores *logit* estimados.

Avaliação do modelo final

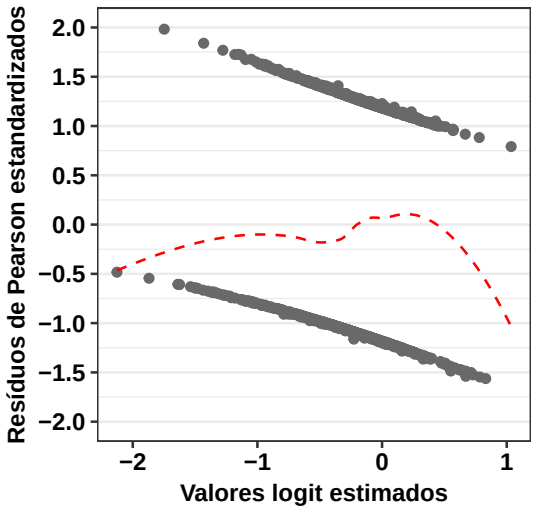
Para avaliar a qualidade de ajustamento do modelo foi aplicado o teste da soma não ponderada de quadrados, onde se obteve um valor-p de 0.90. Como tal, não existe evidência para rejeitar a hipótese nula de que as verdadeiras probabilidades são aquelas especificadas pelo modelo.

Prosseguiu-se então para o diagnóstico do modelo para confirmar a adequabilidade do modelo ajustado. Pela análise do gráfico da esquerda da figura 5.13, verificou-se que um resíduo estandardizado teve um valor próximo a 2, sendo considerado como potencial *outlier*. Observou-se também que a curva *lowess* incorporada no gráfico resultou aproximadamente numa linha horizontal com interceção em 0, exceto nos valores mais extremos de *logit*, o que sugere que o modelo está correto e não existem *outliers* significativos.

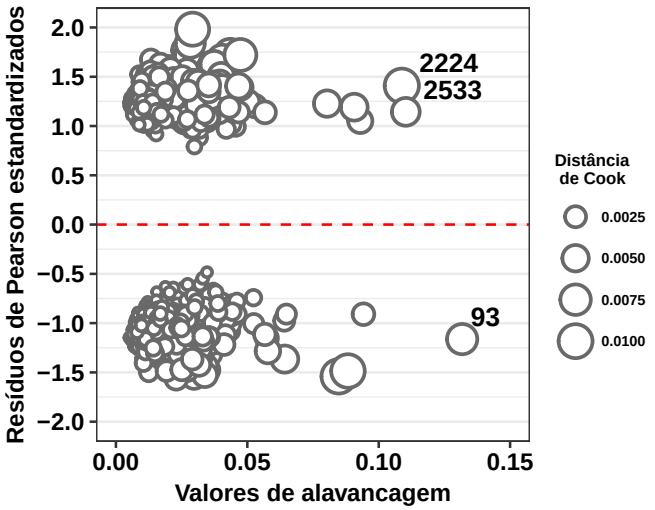
Na tabela 5.10 são apresentados os valores dos resíduos de Pearson estandardizados (prs_i), de alavacagem (h_{ii}) e distância de Cook ($\Delta\beta_i$) que verificaram maior magnitude no conjunto de dados. Os participantes com ID 2066 e 1726 apresentaram os resíduos Pearson estandardizados com maior valor absoluto (1.997 e 1.847), contudo os seus valores de alavancagem foram inferiores ao ponto crítico de 0.0394 $(2(18+1)/963)$, pelo que se decidiu mantê-los no ajustamento do modelo. O indivíduo com ID 358 apresentou igualmente um valor de alavancagem abaixo do ponto crítico e uma baixa distância de Cook, tendo assim pouca influência no ajustamento do modelo, pelo que se decidiu manter essa observação. Uma vez que os participantes de ID 2224, 2533 e 93 apresentaram maiores valores de alavancagem (por ordem crescente) e que a observação de ID 1353 apresentou valores elevados relativamente às restantes, tanto de alavancagem como de distância de Cook , essas observações foram retiradas individualmente do conjunto de dados para a comparação dos coeficientes do modelo reajustado com os do *Modelo II*. A eliminação de cada uma dessas observações provocou alterações acima de 10% em determinados coeficientes, maioritariamente associados às variáveis `pc3`, `idade` ou `cia_energia_kcal`. No entanto, estas variáveis não são significativas no *Modelo II* e nos novos modelos reajustados, pelo que não se justifica a eliminação dessas observações.

Tabela 5.10: Valores dos resíduos de Pearson estandardizados (prs_i), de alavacagem (h_{ii}) e distância de Cook ($\Delta\beta_i$)

ID	rps_i	h_{ii}	$\Delta\beta_i$
93	-1.137400	0.13175116	0.007334774
358	1.773023	0.02651145	0.005279724
710	-1.533016	0.08478721	0.010429508
1353	-1.476851	0.08829610	0.009705556
1726	1.847304	0.02828453	0.006594869
2066	1.997187	0.02917747	0.009353087
2224	1.393867	0.10878174	0.010251074
2533	1.123686	0.11030846	0.005796054



(a) Gráfico dos resíduos de Pearson estandardizados em função dos valores *logit* estimados. A linha tracejada vermelha representa a curva *lowess smooth* dos resíduos contra as probabilidades logísticas estimadas.



(b) Gráfico de influência proporcional dos resíduos de Pearson estandardizados em função dos valores de alavancagem. O tamanho dos círculos é proporcional à distância de Cook apresentada para as respetivas observações.

Figura 5.13: Gráficos utilizados no diagnóstico do *Modelo II*

6 Discussão

Neste estudo foram identificados três padrões alimentares principais na população estudada da Lousã com recurso à ACP: um padrão tradicional português, um padrão saudável e um padrão menos saudável, designado por "Petiscos". Foram encontradas associações entre a DMI e o padrão "Saudável"e "Dieta tradicional portuguesa"controlando para as variáveis relativas à idade, exercício, sexo, perímetro abdominal, dislipidemia, álcool e hábitos tabágicos. Porém estas associações foram estatisticamente significativas apenas ao nível de $\alpha = 0.1$. O padrão "Saudável"demonstrou ser benéfico relativamente à doença, enquanto que o padrão "Dieta tradicional portuguesa"revelou propiciar o seu risco. Não se observou qualquer associação entre a doença e o padrão "Petiscos", onde não só o coeficiente estimado associado ao *score* para esse padrão revelou ser próximo de zero, como valor-p para o teste de Wald foi aproximadamente 1.

Na área da epidemiologia nutricional tem-se verificado uma crescente utilização de métodos destinados a determinar padrões alimentares. Entre esses métodos, a análise em componentes principais e a análise classificatória são os mais frequentemente utilizados (Newby and Tucker, 2004). A variância cumulativa dos três padrões alimentares obtidos com ACP foi de apenas 23.13%, que apesar de baixa, apresentou ser da mesma ordem de grandeza que noutros estudos que extraíram padrões alimentares por ACP (Chiu et al., 2014; Islam et al., 2014). Estes padrões foram, de maneira geral, consistentes com os *clusters* obtidos por AC, apesar de algumas diferenças, confirmando os resultados obtidos com a ACP.

O padrão "Dieta tradicional portuguesa"refletiu uma dieta típica da região interior de Portugal, caracterizado por elevados consumos de carnes vermelhas, bacalhau, acompanhamentos (arroz, massa, batatas cozidas ou assadas), pão branco ou integral, tostas e broa, azeite, bebidas alcoólicas (vinho, cerveja e bebidas brancas) e café. Este padrão apresentou também um consumo reduzido de cereais e iogurte. Estudos realizados noutras regiões geográficas identificaram igualmente padrões típicos da própria cultura gastronómica (Randall et al., 1992; Velie et al., 2005). O padrão "Saudável"foi caracterizado por elevados consumos de iogurte, peixe (numa aparente substituição das carnes vermelhas e brancas) e de fruta, hortícolas, salada e sopa. Este padrão apresentou ser semelhante a outros descritos na literatura relativamente a outras populações, com a designação de “Prudent” ou “Healthy” (Fung et al., 2001; Hu et al., 2000; Slattery et al., 1998; Magalhães et al., 2012), partilhando muitas das características da dieta mediterrânica tradicional (Willett et al., 1995; Trichopoulou et al., 1995; Hu et al., 1999). Finalmente, o padrão "Petiscos"foi caracterizado por uma ingestão elevada de *snacks*, refrigerantes, doces, batatas fritas, marisco, e por um baixo consumo de peixe. Este padrão foi similar a um dos padrões identificados num estudo envolvendo participantes com idade igual ou superior a 60 anos residentes em Botucatu, uma região interior do Brasil, denominado de “Snacks and weekend meal”

(Ferreira et al., 2014).

Concluiu-se então que os padrões encontrados foram consistentes com outros encontrados na literatura, no entanto, o número de padrões e parte das componentes alimentares associadas a cada padrão diferiram de outros estudos (Chiu et al., 2014; Islam et al., 2014; Martínez-González et al., 2011; Fung et al., 2001; Hu et al., 2000; Fung et al., 2004). Os diferentes resultados obtidos poderão resultar do número e tipo de alimentos disponíveis, e respetivo agrupamento para a análise em componentes principais; e também das preferências alimentares individuais, dependentes de fatores culturais, sociais, ambientais, económicos e de estilo de vida.

A má e pobre alimentação da amostra da Lousã refletiu-se no decréscimo substancial de adesão a cada padrão alimentar com o aumento da idade. No entanto, não pareceram existir diferenças consideráveis entre os padrões relativamente à idade. Os indivíduos que com maior adesão ao padrão "Dieta tradicional portuguesa"tenderam a ser do sexo masculino, maiores consumidores de álcool, a serem fumadores ou ex-fumadores e a consumir maior número de pacotes de tabaco anuais. Entre os homens, os valores do perímetro abdominal e de IMC tenderam a ser menores para o padrão "Saudável", apesar das diferenças entre padrões não serem de grande magnitude. Já para as mulheres, apenas os valores de IMC médio registaram diferenças entre os padrões (novamente de baixa magnitude), onde o padrão "Dieta tradicional portuguesa"pareceu ter valores maiores. Os indivíduos do estudo, de modo geral, têm de 1 a 6 anos de escolaridade, não tendo sido registadas diferenças entre os vários padrões. Relativamente às restantes características, não pareceram existir diferenças consideráveis entre os vários padrões alimentares.

A análise da relação entre padrões alimentares e doenças oculares é uma abordagem relativamente nova na área da epidemiologia oftalmológica, onde apenas três estudos se encontram publicados: um deles com o objetivo de analisar a associação dos padrões com o calibre dos vasos retinianos (McEvoy et al., 2013) e os outros dois com o risco de DMI (Chiu et al., 2014; Islam et al., 2014). A utilização de padrões alimentares oferece inúmeras vantagens relativamente aos estudos que se focam apenas em nutrientes e alimentos a nível individual uma vez que tem em consideração a dieta como um todo, refletindo mais proximamente o mundo real, onde qualquer efeito sinergístico entre os nutrientes ou alimentos poderá ser mais facilmente detetado e o resultado mais compreensivamente traduzido para um aconselhamento nutricional.

Este estudo foi o primeiro a analisar a relação entre a DMI e padrões alimentares derivados por métodos *a posteriori* para uma população portuguesa. As associações entre os vários padrões alimentares e o risco da doença foram, de modo geral, nas direções esperadas, demonstrando também consistência com as associações entre o consumo de nutrientes e alimentos observadas em estudos epidemiológicos prévios (Chiu et al., 2014; Islam et al., 2014).

O consumo de alimentos positivamente correlacionados com o padrão "Saudável", como as frutas, vegetais, peixe e alimentos de baixo índice glicémico demonstraram ter um efeito benéfico relativamente à doença (Swenor et al., 2010; Tan et al., 2009) e consistem em alguns dos principais figurantes da típica dieta mediterrânica (Costacou et al., 2003). Vários estudos têm utilizado diferentes índices construídos com base nessa dieta, como o estudo *Carotenoids Age-Related Eye Disease Study* (CAREDS) (Mares et al., 2011), que usou o *2005 Healthy Eating Index* modificado (mHEI) e o estudo realizado por Merle e colaboradores (Merle et al., 2015), que fez uso do *score* alternativo da dieta mediterrânica (aMeDi) para estimar a adesão dos participantes a esse modelo de consumo alimentar. Ambos os estudos reportaram

uma associação entre uma alta adesão à dieta mediterrânica e um menor risco de DMI precoce ou de progressão para uma forma tardia da doença, respetivamente. Numa região costeira de Portugal (Mira) foi igualmente reportada uma diminuição do risco de DMI com a maior adesão ao *score* da dieta mediterrânica (Farinha et al., 2015).

Relativamente aos dois estudos que avaliaram a associação entre os padrões alimentares empíricos e o risco de DMI, no estudo *The Relationship of Major American Dietary Patterns to Age-Related Macular Degeneration* foi derivado um padrão designado de "Oriental", caracterizado por um elevado consumo de vegetais, legumes, fruta, cereais integrais, tomates e marisco, muito semelhante ao padrão "Saudável"obtido neste trabalho (Chiu et al., 2014). O padrão "Oriental"revelou ser benéfico tanto para a forma precoce, como para as formas tardias de DMI. Já no estudo *Dietary Patterns and Their Associations with Age-Related Macular Degeneration: The Melbourne Collaborative Cohort Study*, os resultados sugeriram também um efeito protetor relativamente à doença de um padrão caracterizado por peixe, frango, muesli e uma diversidade de vegetais, e por baixo consumo de pão branco. No entanto, esse efeito foi verificado apenas para as formas tardias de DMI (Islam et al., 2014).

Por sua vez, os alimentos que caracterizam o padrão "Saudável"são ricos em diversos nutrientes associados a um menor risco ou progressão de DMI, como as vitaminas C e E, o zinco, e outros nutrientes cujas quantidades de consumo se encontram em elevados níveis nas fontes alimentares aqui avaliadas (INSA, 2006). Entre esses nutrientes encontram-se dois ácidos gordos polinsaturados ómega-3, o ácido docosa-hexaenoico (DHA) e o ácido eicosapentaenoico (EPA), abundantes no peixe (Swenor et al., 2010; Tan et al., 2009); e os carotenóides luteína e zeaxantina, abundantes em diversos produtos vegetais como frutas, especiarias, alface e vegetais de folha escura. Deste modo, a associação negativa (protetora) entre o padrão alimentar "Saudável"e o risco de DMI poderá resultar destes alimentos.

O elevado consumo de carnes vermelhas, álcool e alimentos de baixo índice glicémico, como arroz, massa, batatas e pão branco ou tostas, associados ao padrão "Dieta tradicional portuguesa", demonstrou ter um efeito prejudicial relativamente à doença (Chong et al., 2009, 2008; Adams et al., 2012; Chiu et al., 2007a,b). Deste modo, os resultados desses estudos, que analisaram os alimentos a nível individual, foram consistentes com os aqui obtidos, onde se observou uma associação positiva entre o padrão "Dieta tradicional portuguesa"e o risco de DMI. Este padrão é aquele que partilha mais características com o padrão "Red meat"identificado pelos autores (Islam et al., 2014), que concluíram que o baixo consumo de carnes vermelhas poderia estar associado com um menor risco de DMI. Comparativamente com os outros padrões, a maior adesão ao padrão tradicional tendeu para um menor consumo de antioxidantes, cálcio, zinco e, ao contrário do esperado, menores níveis de hidratos de carbono simples e de total de hidratos de carbono disponíveis e também de gorduras (excepto ácidos gordos monoinsaturados e colesterol). Tendeu também para um maior consumo de ferro, colesterol e proteínas.

O padrão "Petiscos"não se encontrou associado com o risco de DMI, apesar de ser caracterizado por um elevado consumo de alimentos ricos em gorduras *trans*, carnes processadas e alimentos de alto índice glicémico, que têm sido reportados como tendo uma associação positiva com o risco de DMI (Chiu et al., 2007a,b; Islam et al., 2014; Chiu et al., 2014). Tal poderá explicar-se pelo facto de os padrões identificados não serem correlacionados, onde um indivíduo com alto *score* no padrão "Petiscos"não tem necessariamente de apresentar um baixo *score* noutro padrão, podendo os constituintes dietéticos dos diversos alimentos interagir e afetar o risco da doença. De facto, na análise classificatória foi obtido

um *cluster* denominado de "Dieta equilibrada", onde os indivíduos incluídos nesse grupo praticaram uma dieta composta por alimentos do padrão "Saudável"e do padrão "Petiscos". Deste modo, os efeitos prejudiciais associados aos alimentos que caracterizam o padrão "Petiscos", poderão estar a ser equilibrados pelos efeitos protetores associados ao padrão "Saudável".

Tal como em todos os estudos observacionais, a natureza transversal deste estudo limita a sua força na definição de causalidade e na capacidade de fazer recomendações dietéticas. O confundimento residual poderá ainda ser uma preocupação, uma vez que os padrões dietéticos poderão ser apenas uma componente de um estilo de vida que, em geral, é responsável pela relação adjacente com a doença. Inclusivamente, no estudo (Nonyane et al., 2010) foi sugerido que diferenças na prevalência da forma tardia de DMI em várias etnias e regiões geográficas possam refletir variações genéticas (Nonyane et al., 2010). O polimorfismo de um nucleótido Y402H, no gene fator H do complemento (CFH), foi sugerido por Nonyane e colegas como estando implicado na DMI. O alelo de risco aparenta prejudicar a função reguladora do gene, levando a uma sobreativação do complemento e, assim, aumentando o risco da doença (Schramm et al., 2014). As populações de descendência europeia apresentaram maiores frequências de alelos de risco (em ambos) e maior prevalência de DMI tardia do que os descendentes chineses e japoneses (Nonyane et al., 2010). Adicionalmente, os resultados do estudo de Merle e colegas sugeriram que uma maior adesão a uma dieta mediterrânica estava associada a um risco reduzido de progressão de DMI e que essa associação poderia ser modelada pela variante do CFH Y402H (Merle et al., 2015). Dado que não foi realizada a recolha de dados sobre a história familiar de DMI nem a genotipagem dos doentes, potenciais interações entre o perfil genético e a alimentação praticada pelos participantes não foram analisadas. Deste modo, essas potenciais interações poderão estar a exercer influência sobre a magnitude das associações observadas entre os padrões alimentares aqui obtidos e o risco de DMI.

É também importante ter em conta que os perfis dos padrões podem mudar ao longo do tempo, dada a mudança nas preferências alimentares e disponibilidade dos alimentos, ou até mesmo devido ao diagnóstico de determinadas patologias, como diabetes e participantes com histórico de ataque cardíaco ou angina (Islam et al., 2014), que resultam numa potencial alteração dos hábitos alimentares. Ao contrário de alguns estudos (Islam et al., 2014; Chiu et al., 2014), os participantes diabéticos não foram excluídos do estudo por decisão do investigador principal, devido a razões por si apresentadas e já anteriormente enumeradas, que justificam a ausência de viés ao nível dos padrões de consumo alimentares aqui obtidos. Inclusivamente, não foi possível realizar o ajustamento para potencial suplementação do AREDS e para a situação económica dos participantes, uma vez que a informação recolhida para estas variáveis foi insuficiente ou com pouca especificação.

O subdiagnóstico de DMI poderá ter ocorrido, uma vez que o diagnóstico de DMI foi baseado apenas em fotografias do fundo ocular, não havendo dados de tomografia de coerência ótica ou de angiografia por fluoresceína. Inclusivamente, não se fez aqui distinção das várias formas da doença, assim como nos outros estudos que estudaram a associação entre nutrientes, alimentos ou padrões com DMI (Islam et al., 2014; Chiu et al., 2014; Merle et al., 2015). Tal deveu-se ao facto de existirem poucos participantes com formas avançadas de DMI ($n = 39$, todos foram incluídos nas análises estatísticas efetuadas, correspondendo a 9.11% dos participantes com DMI incluídos nas análises. A observação com ID 2443 correspondeu a um participante com DMI avançada, e não foi incluída no modelo final de regressão logística binária construído), não perfazendo uma dimensão adequada para se realizarem estudos separados para as formas precoces e avançadas da doença. Como já mencionado, o estudo (Islam et al., 2014) reportou associações entre alguns dos padrões identificados com a DMI tardia, no

entanto, os autores não observaram quaisquer associações com a forma precoce da doença. Nesse estudo foi também sugerido que tal poderia ter ocorrido devido a uma menor relação da dieta com a prevenção primária da doença (prevenção do desenvolvimento de DMI precoce) comparativamente à prevenção secundária (prevenção da progressão da forma precoce para formas avançadas da doença) (Chong et al., 2007). Consequentemente, a magnitude das associações aqui encontradas poderá ter sido influenciada em diferentes extensões pelos diferentes estadios da doença.

Questões metodológicas

O uso de questionários de frequência alimentar apresenta limitações na medida de quantidades absolutas, a exemplo tem-se a tendência à subestimação do consumo alimentar como consequência do consumo de pratos mistos cozinhados com diferentes ingredientes individuais e temperos, que contribuem altamente para o consumo de nutrientes, mas não são considerados nos cálculos dos consumos alimentares (Yun et al., 2013; Ahn et al., 2004; Shim et al., 1997; Yun et al., 2009). Deste modo, foram eliminados os participantes nos extremos 1 % de consumos energéticos altos e baixos (Islam et al., 2014). Salienta-se também o facto de ser improvável que os participantes que verificam estes consumos energéticos extremos tenham reportado um consumo usual real, logo a sua inclusão poderia resultar na distorção dos padrões alimentares identificados. Um potencial viés de seleção resultante da exclusão desses participantes foi possivelmente pequeno, dado que não existiram diferenças significativas na prevalência de DMI e nas características potencialmente relevantes para o risco da doença entre as pessoas incluídas e excluídas do estudo. Outras limitações adjacentes a este tipo de questionário resultam particularmente das restrições impostas por uma lista fixa de alimentos, do recurso à memória, da percepção das porções médias e da interpretação das questões (Lopes et al., 2006). Para minimizar estes potenciais erros, a administração do questionário foi realizada através de uma entrevista pessoal por um entrevistador treinado, o que permitiu uma melhor assistência ao participante, o esclarecimento de dúvidas no momento e a deteção de algumas contradições de resposta. O treino rigoroso dos inquiridores e uniformizaçao periódica dos procedimentos permitiu também a minimização do viés associado às diferenças de atuação de na recolha de informação por parte do inquiridor e o viés associado ao prestígio social, introduzido pela presença do entrevistador.

Outra fonte de erro que poderá contribuir para a diminuição da validade dos questionários de frequência alimentar é a ampla variação intra-pessoal do tamanho das porções de alimentos ingeridas, embora a frequência seja referida como o determinante mais importante no cálculo do consumo alimentar (Lopes et al., 2006). A especificação das porções médias padrão adaptadas à população em estudo aumenta a objetividade das questões, a rapidez de recolha e subsequente análise de dados, diminui os custos e a complexidade do questionário e não introduz um erro significativo na estimativa da ingestão de alimentos e nutrientes, razões que justificam a opção por esta alternativa (Lopes et al., 2006). A falta de informação detalhada sobre os alimentos inerentes a este tipo de questionário poderia levar a alguma falta de precisão, no entanto está descrito que o custo adicional de questões abertas sobre os alimentos contribui apenas para um pequeno incremento na estimativa do nutriente (Willett, 1998). Além disso, a composição de um alimento varia nas diferentes áreas de um país e com as diferentes estações do ano, e, como tal, existe uma natural e inevitável diminuição da precisão do consumo alimentar (Whiting and Leverton, 1960). Note-se inclusivamente que a dieta foi avaliada apenas por uma única administração do questionário, podendo este não ser representativo do consumo ao longo da vida do participante. Os erros aleatórios na medida do consumo alimentar poderão também ter tido impacto nas associações encontradas neste projeto.

A ACP e a AC são técnicas úteis para a avaliação dos padrões alimentares, e o máximo de informação pode ser obtido quando estes dois métodos são utilizados. Os pontos fortes e fracos destas técnicas encontram-se descritos em detalhe por Michels and Schulze (2015). Vários autores citam a subjetividade destas técnicas como a sua principal desvantagem, devido às várias decisões a serem determinadas *a priori* pelo investigador, como por exemplo, o agrupamento dos itens alimentares do questionário e formato das variáveis a serem utilizadas na análise, o número de padrões a reter, ou a atribuição de designações a cada padrão. No corrente estudo, vários passos foram tomados para diminuir essa subjetividade, no entanto, não foi aqui avaliada a validade e reprodutibilidade dos padrões alimentares estimados a partir do questionário de frequência alimentar aplicado aos participantes.

O ajustamento para a energia dos alimentos ou grupos de alimentos prévio à ACP não foi realizado, existindo entre vários autores a preocupação de que qualquer associação encontrada entre uma ocorrência de interesse e um padrão alimentar que represente uma dieta rica em alimentos de elevada densidade energética poder não ser devida a um efeito real dos alimentos em si, mas a uma associação com o consumo total energético (Willett et al., 1997; Northstone et al., 2008). No entanto, no estudo *Adjusting for energy intake in dietary pattern investigations using principal components analysis* foram avaliadas as diferenças entre os padrões obtidos por ACP com base em dados ajustados e não ajustados para a energia, concluindo-se que os padrões obtidos eram comparáveis e robustos ao ajustamento por energia e recomendando-se apenas o ajustamento para a energia numa fase mais tardia do processo analítico (Northstone et al., 2008). Neste trabalho, o consumo energético foi controlado no modelo de regressão binária, de modo a que os seus efeitos reais no risco de DMI pudessem ser claramente determinados.

Na área da investigação epidemiológica causal, os modelos de regressão múltipla são frequentemente utilizados para controlar o confundimento. No entanto, a modelação das variáveis de confundimento contínuas não tem recebido muita atenção. Uma revisão da prática atual na aplicação do ajustamento para confundidores nos estudos epidemiológicos observacionais revelou que a relação funcional entre as variáveis de confundimento contínuas e a ocorrência de interesse é raramente reportada (Groenwold et al., 2013). Inclusivamente, os únicos dois estudos que analisaram a associação entre padrões alimentares e a prevalência e/ou progressão de DMI limitaram-se ao pressuposto de linearidade das variáveis de confundimento contínuas e/ou à respetiva dicotomização ou estratificação segundo valores frequentemente utilizados na literatura (Chiu et al., 2014; Islam et al., 2014). A não observação da eventual não linearidade para essas variáveis poderá ter resultado num confundimento residual considerável e, consequentemente, à distorção do verdadeiro efeito das variáveis de interesse, ou seja, dos padrões alimentares no risco da doença.

Neste trabalho procurou-se minimizar o confundimento residual através do desenvolvimento de uma estratégia que permitisse simultaneamente a seleção de variáveis e a seleção das formas funcionais das variáveis contínuas, com base nos métodos e resultados de outros artigos (Binder et al., 2013; Benedetti and Abrahamowicz, 2004). Para as variáveis contínuas, incluindo as variáveis de interesse, foi testada a não linearidade dos modelos com as formas polinomiais (quadrática e cúbica) e do modelo resultante da aplicação da técnica de *restricted cubic splines*, onde foram obtidas componentes RCS para essas variáveis, após determinado o número de nós ótimo pelo critério de informação AIC.

As variáveis de confundimento `jafumador`, `pacotesano` e `alcool` foram as únicas para as quais se observou significância estatística ao nível $\alpha = 0.05$. As variáveis `pacotesano` e `alcool` foram também

as únicas que apresentaram não linearidade ao nível $\alpha = 0.05$, a primeira sendo incluída no modelo com uma componente RCS de 5 nós e a segunda com uma componente RCS de 3 nós. A interpretação dos coeficientes associados às componentes destas variáveis é, como se pode perceber, difícil e pouco intuitiva, salientando uma das principais limitações da utilização deste tipo de metodologia estatística. Além disso, a análise gráfica (dados aqui não demonstrados) dos *logit* estimados ao longo do domínio dessas variáveis não permitiu uma interpretação clínica clara. No entanto, a aplicação duma modelação mais flexível a estas variáveis permite um controlo mais completo do confundimento e, assim, uma estimação mais correta e precisa da magnitude da relação entre os padrões alimentares e a doença (Brenner and Blettner, 1997).

7 Conclusão

Neste estudo foram identificados três padrões alimentares com ACP numa subpopulação rural do centro de Portugal com idade superior a 55 anos e foram analisadas as associações dos padrões obtidos com o risco de desenvolvimento de DMI. Os dados aqui apresentados suportam a opinião de que a dieta tem um papel importante no desenvolvimento de DMI. Uma dieta saudável, rica em legumes, saladas, frutas e peixe demonstrou ser protetora da doença, enquanto que uma dieta tradicional à base de carnes vermelhas, alimentos de alto índice glicémico como as batatas, arroz e massa, farináceos e com elevado consumo de álcool demonstrou potenciar o risco da doença.

Ao conhecimento atual, este foi o primeiro estudo a investigar a associação entre padrões alimentares e o risco de DMI numa subpopulação portuguesa. Neste estudo procurou-se também minimizar o confundimento residual no processo de modelação através do desenvolvimento e aplicação de uma estratégia de seleção simultânea de variáveis de confundimento e da forma funcional das variáveis contínuas, com vista a diminuir a distorção da verdadeira associação entre os padrões alimentares e o risco de DMI.

Salienta-se que os resultados da análise de padrões alimentares não têm a capacidade de revelar a quantidade absoluta de determinados alimentos a consumir ou a evitar, contudo oferecem uma indicação de quais os padrões alimentares potencialmente importantes a considerar para prevenir o desenvolvimento de DMI. É também importante ter em consideração que a dieta é apenas uma componente de um estilo de vida que, em geral, é responsável pela relação adjacente com a DMI. Como tal, a divulgação e implementação de estratégias que promovam comportamentos coletivos saudáveis poderá não só ter um impacto benéfico na prevenção de DMI, como também na prevenção de outras patologias crónicas, com a consequente melhoria da qualidade de vida dos idosos.

Futuramente seria de interesse investigar potenciais interações entre o perfil genético e alimentar dos participantes e também os mecanismos biológicos pelos quais os padrões alimentares aqui obtidos possam estar a influenciar o desenvolvimento e a progressão de DMI, de modo a permitir identificar populações alvo e intervenções dietéticas que se revelem mais eficazes para os indivíduos dessas populações. Uma vez que vários estudos sugeriram que uma menor magnitude da associação da dieta com a forma precoce de DMI, comparativamente às formas avançadas, seria relevante realizar o estudo separado da associação dos padrões alimentares com o risco de DMI na forma precoce e com o risco de progressão da doença para formas avançadas, para a melhor compreensão do potencial papel da dieta nas várias fases de prevenção da doença.

Bibliografia

Abdi, H. (2003). *Encyclopedia for research methods for the social sciences*, chapter Factor rotations in factor analyses, pages 792–795. Thousand Oaks (CA): Sage.

Adams, M. K., Chong, E. W., Williamson, E., Aung, K. Z., Makeyeva, G. A., Giles, G. G., English, D. R., Hopper, J., Guymer, R. H., Baird, P. N., et al. (2012). 20/20–Alcohol and Age-related Macular Degeneration The Melbourne Collaborative Cohort Study. *American Journal of Epidemiology*, 176(4):289–298.

Age-Related Eye Disease Study Research Group et al. (2000). Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: Age-Related Eye Disease Study Report Number 3. *Ophthalmology*, 107(12):2224–2232.

Age-Related Eye Disease Study Research Group et al. (2008). A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8. *Archives of Ophthalmology*, 126(9):1251.

Agresti, A. (2012). *Categorical Data Analysis*. Wiley, Hoboken, NJ, 3 edition.

Ahn, Y., Lee, J. E., Cho, N. H., Shin, C., Park, C., Oh, B. S., and Kimm, K. (2004). Validation and calibration of semi-quantitative food frequency questionnaire: with participants of the Korean Health and Genome Study. *Korean Journal of Community Nutrition*, 9(2):173–182.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Allison, P. D. (2014). Measures of fit for logistic regression. In *SAS Global Forum, Washington, DC*.

Arnarsson, A., Sverrisson, T., Stefánsson, E., Sigurdsson, H., Sasaki, H., Sasaki, K., and Jonasson, F. (2006). Risk factors for five-year incident age-related macular degeneration: the Reykjavik Eye Study. *American Journal of Ophthalmology*, 142(3):419–428.

Auguie, B. and Antonov, A. (2016). Miscellaneous functions for grid graphics. R package version 2.2.1.

Bagheri, A., Midi, H., and Imon, A. (2010). The effect of collinearity-influential observations on collinear data set: A monte carlo simulation study. *Journal of Applied Sciences*, 10(18):2086–2093.

Baik, I., Ascherio, A., Rimm, E. B., Giovannucci, E., Spiegelman, D., Stampfer, M. J., and Willett, W. C. (2000). Adiposity and Mortality in Men. *American Journal of Epidemiology*, 152(3):264–271.

Beatty, S., Koh, H.-H., Phil, M., Henson, D., and Boulton, M. (2000). The Role of Oxidative Stress in the Pathogenesis of Age-Related Macular Degeneration. *Survey of Ophthalmology*, 45(2):115–134.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2013). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley-Interscience, New York, NY.

Bender, R. (2009). Introduction to the use of regression models in epidemiology. *Cancer Epidemiology*, 471:179–195.

Benedetti, A. and Abrahamowicz, M. (2004). Using generalized additive models to reduce residual confounding. *Statistics in Medicine*, 23(24):3781–3801.

Bibby, J., Kent, J., and Mardia, K. (1980). Multivariate Analysis.

Binder, C. J., Chang, M.-K., Shaw, P. X., Miller, Y. I., Hartvigsen, K., Dewan, A., and Witztum, J. L. (2002). Innate and acquired immunity in atherogenesis. *Nature Medicine*, 8(11):1218–1226.

Binder, H., Sauerbrei, W., and Royston, P. (2013). Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine*, 32(13):2262–2277.

Bird, A., Bressler, N., Bressler, S., Chisholm, I., Coscas, G., Davis, M., De Jong, P., Klaver, C., Klein, B., Klein, R., et al. (1995). An International Classification and Grading System for Age-related Maculopathy and Age-related Macular Degeneration. *Survey of ophthalmology*, 39(5):367–374.

Blumenkranz, M. S., Russell, S. R., Robey, M. G., Kott-Blumenkranz, R., and Penneys, N. (1986). Risk Factors in Age-related Maculopathy Complicated by Choroidal Neovascularization. *Ophthalmology*, 93(5):552–558.

Borger, P. H., van Leeuwen, R., Hulsman, C. A., Wolfs, R. C., van der Kuip, D. A., Hofman, A., and de Jong, P. T. (2003). Is there a direct association between age-related eye diseases and mortality?: The Rotterdam Study. *Ophthalmology*, 110(7):1292–1296.

Brenner, H. and Blettner, M. (1997). Controlling for Continuous Confounders in Epidemiologic Research. *Epidemiology*, 8(4):429–434.

Brewer, G. J. (2007). Iron and copper toxicity in diseases of aging, particularly atherosclerosis and Alzheimer's disease. *Experimental Biology and Medicine*, 232(2):323–335.

Buch, H. (2005). Fourteen-year incidence of age-related maculopathy and cause-specific prevalence of visual impairment and blindness in a Caucasian population: the Copenhagen City Eye Study. *Acta Ophthalmologica Scandinavica*, 83(thesis):5–32.

Cachulo, M. d. L., Lobo, C., Figueira, J., Ribeiro, L., Laíns, I., Vieira, A., Nunes, S., Costa, M., Simão, S., Rodrigues, V., et al. (2015). Prevalence of Age-Related Macular Degeneration in Portugal: The Coimbra Eye Study-Report 1. *Ophthalmologica*, 233(3-4):119–127.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Chen, X., Rong, S. S., Xu, Q., Tang, F. Y., Liu, Y., Gu, H., Tam, P. O., Chen, L. J., Brelén, M. E., Pang, C. P., et al. (2014). Diabetes Mellitus and Risk of Age-Related Macular Degeneration: A Systematic Review and Meta-Analysis. *PloS ONE*, 9(9):108–196.

Chiu, C., Klein, R., Milton, R., Gensler, G., and Taylor, A. (2009). Does eating particular diets alter the risk of age-related macular degeneration in users of the Age-Related Eye Disease Study supplements? *British Journal of Ophthalmology*, 93(9):1241–1246.

Chiu, C.-J., Chang, M.-L., Zhang, F. F., Li, T., Gensler, G., Schleicher, M., and Taylor, A. (2014). The Relationship of Major American Dietary Patterns to Age-Related Macular Degeneration. *American Journal of Ophthalmology*, 158(1):118–127.

Chiu, C.-J., Milton, R. C., Gensler, G., and Taylor, A. (2007a). Association between dietary glycemic index and age-related macular degeneration in nondiabetic participants in the Age-Related Eye Disease Study. *The American Journal of Clinical Nutrition*, 86(1):180–188.

Chiu, C.-J., Milton, R. C., Klein, R., Gensler, G., and Taylor, A. (2007b). Dietary carbohydrate and the progression of age-related macular degeneration: a prospective study from the Age-Related Eye Disease Study. *The American Journal of Clinical Nutrition*, 86(4):1210–1218.

Chiu, C.-J. and Taylor, A. (2007). Nutritional antioxidants and age-related cataract and maculopathy. *Experimental Eye Research*, 84(2):229–245.

Cho, E., Hankinson, S. E., Willett, W. C., Stampfer, M. J., Spiegelman, D., Speizer, F. E., Rimm, E. B., and Seddon, J. M. (2000). Prospective Study of Alcohol Consumption and the Risk of Age-Related Macular Degeneration. *Archives of Ophthalmology*, 118(5):681–688.

Chong, E. W., Wong, T. Y., Kreis, A. J., Simpson, J. A., and Guymer, R. H. (2007). Dietary antioxidants and primary prevention of age related macular degeneration: systematic review and meta-analysis. *BMJ*, 335(7623):755.

Chong, E. W.-T., Kreis, A. J., Wong, T. Y., Simpson, J. A., and Guymer, R. H. (2008). Alcohol consumption and the risk of age-related macular degeneration: a systematic review and meta-analysis. *American Journal of Ophthalmology*, 145(4):707–715.

Chong, E. W.-T., Simpson, J. A., Robman, L. D., Hodge, A. M., Aung, K. Z., English, D. R., Giles, G. G., and Guymer, R. H. (2009). Red meat and chicken consumption and its association with age-related macular degeneration. *American Journal of Epidemiology*, 169(7):867–876.

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., and Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British journal of health psychology*, 10(3):329–358.

Clemons, T., Rankin, M., and McBee, W. (2006). Cognitive impairment in the Age-Related Eye Disease Study: AREDS report No. 16. *Archives of Ophthalmology*, 124(4):537–543.

Cleveland, W. S. and Loader, C. (1996). Smoothing by local regression: Principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*, pages 10–49. Physica-Verlag HD. DOI: 10.1007/978-3-642-48425-4_2.

Copas, J. (1989). Unweighted Sum of Squares Test for Proportions. *Applied Statistics*, 38(1):71–80.

Correia, J. (2002). Etiopatogenia da DMI. *Acta Oftalmológica*, 12:23–34.

Costacou, T., Bamia, C., Ferrari, P., Riboli, E., Trichopoulos, D., and Trichopoulou, A. (2003). Tracing the Mediterranean diet through principal components and cluster analyses in the Greek population. *European Journal of Clinical Nutrition*, 57(11):1378–1385.

Curcio, C. A., Johnson, M., Huang, J.-D., and Rudolf, M. (2009). Aging, age-related macular degeneration, and the response-to-retention of apolipoprotein B-containing lipoproteins. *Progress in Retinal and Eye Research*, 28(6):393–422.

De Boor, C. (1978). *A Practical Guide to Splines*, volume 27. Springer, New York.

Delaney Jr, W. and Oates, R. (1982). Senile macular degeneration: a preliminary study. *Annals of Ophthalmology*, 14(1):21.

Delcourt, C., Michel, F., Colvez, A., Lacroux, A., Delage, M., and Vernet, M.-H. (2001). Associations of cardiovascular disease and its risk factors with age-related macular degeneration: the POLA study. *Ophthalmic Epidemiology*, 8(4):237–249.

Devlin, U. M., McNulty, B. A., Nugent, A. P., and Gibney, M. J. (2012). The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting. *Proceedings of the Nutrition Society*, 71(4):599–609.

Direção-Geral da Saúde (2005). Programa Nacional de Combate à Obesidade. Lisboa.

Direção-Geral da Saúde (2016). Programa Nacional para a Saúde da Visão – Revisão e Extensão 2020. Lisboa.

dos Santos Silva, I. (1999). Dealing with confounding in the analysis [chapter 14]. In *Cancer Epidemiology: Principles and Methods*, pages 305–333. World Health Organization, Lyon, 2nd revised edition.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York, 1st edition.

Ederer, F., Church, T. R., and Mandel, J. S. (1993). Sample sizes for prevention trials have been too small. *American Journal of Epidemiology*, 137(7):787–796.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). Hierarchical Clustering. In *Cluster Analysis*, pages 71–110. Wiley, Chichester, 5th edition.

Eye Disease Case-Control Study Group et al. (1992). Risk factors for neovascular age-related macular degeneration. The Eye Disease Case-Control Study Group. *Archives of Ophthalmology*, 110(12):1701–1708.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics*, 13(1):342–368.

Farinha, C. L., Mira, F., Santos, L., Nunes, S., Pedroso, A., Laíns, I., Cachulo, M. L., and Silva, R. (2015). Nutritional and lifestyle risk factors in AMD. The Coimbra Eye Study. *Investigative Ophthalmology & Visual Science*, 56(7):3770–3770.

Ferreira, F. A. G., Graça, M. d. S., et al. (1985). Tabela da composição dos alimentos portugueses.

Ferreira, P. M., Papini, S. J., and Corrente, J. E. (2014). Diversity of eating patterns in older adults: A new scenario? *Revista de Nutrição*, 27(1):67–79.

Fox, J. and Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, 87(417):178–183.

Fox, J. and Weisberg, S. (2016). Companion to applied regression. R package v 2.1-4.

Fransen, H. P., May, A. M., Stricker, M. D., Boer, J. M., Hennig, C., Rosseel, Y., Ocké, M. C., Peeters, P. H., and Beulens, J. W. (2014). A posteriori dietary patterns: how many patterns to retain? *The Journal of Nutrition*, 144(8):1274–1282.

Fraser-Bell, S., Wu, J., Klein, R., Azen, S. P., Hooper, C., Foong, A. W., Varma, R., Group, L. A. L. E. S., et al. (2008). Cardiovascular risk factors and age-related macular degeneration: the Los Angeles Latino Eye Study. *American Journal of Ophthalmology*, 145(2):308–316.

Fraser-Bell, S., Wu, J., Klein, R., Azen, S. P., Varma, R., Los Angeles Latino Eye Study Group, et al. (2006). Smoking, alcohol intake, estrogen use, and age-related macular degeneration in Latinos: the Los Angeles Latino Eye Study. *American Journal of Ophthalmology*, 141(1):79–87.

Friedman, D. S., O'Colmain, B. J., Munoz, B., Tomany, S. C., McCarty, C., De Jong, P., Nemesure, B., Mitchell, P., Kempen, J., et al. (2004). Prevalence of age-related macular degeneration in the United States. *Archives of Ophthalmology*, 122(4):564–572.

Fung, T. T., Stampfer, M. J., Manson, J. E., Rexrode, K. M., Willett, W. C., and Hu, F. B. (2004). Prospective study of major dietary patterns and stroke risk in women. *Stroke*, 35(9):2014–2019.

Fung, T. T., Willett, W. C., Stampfer, M. J., Manson, J. E., and Hu, F. B. (2001). Dietary patterns and the risk of coronary heart disease in women. *Archives of Internal Medicine*, 161(15):1857–1862.

Gnanadesikan, R. (2011). *Methods for Statistical Data Analysis of Multivariate Observations*, volume 321. Wiley-Interscience.

Gomes, J. (2011a). Regressão binária – o modelo logístico. Lisboa: DEIO/FCUL.

Gomes, J. (2011b). Regressão linear. Lisboa: DEIO/FCUL.

Gopinath, B., Flood, V. M., Kifley, A., Liew, G., and Mitchell, P. (2015). Smoking, Antioxidant Supplementation and Dietary Intakes among Older Adults with Age-Related Macular Degeneration over 10 Years. *PloS ONE*, 10(3):122–548.

Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, 167(5):523–529.

Groenwold, R. H., Klungel, O. H., Altman, D. G., van der Graaf, Y., Hoes, A. W., and Moons, K. G. (2013). Adjustment for continuous confounders: an example of how to prevent residual confounding. *Canadian Medical Association Journal*, 185(5):401–406.

Gustafsson, K. and Sidenvall, B. (2002). Food-related health perceptions and food habits among older women. *Journal of Advanced Nursing*, 39(2):164–173.

Harrell, F. E. (2014). Regression modeling strategies. *BIOS*, 330.

Harrell Jr, F. (2017). Regression modeling strategies. R package version 5.1-1.

Hawkins, B. S., Bird, A., Klein, R., and West, S. K. (1999). Epidemiology of age-related macular degeneration. *Molecular Vision*, 5(26).

Herne, S. (1995). Research on food choice and nutritional status in elderly people: a review. *British Food Journal*, 97(9):12–29.

Hilbe, J. M. (2009). *Logistic Regression Models*. Chapman and Hall/CRC, Boca Raton, 1st edition.

Hollyfield, J. G., Bonilha, V. L., Rayborn, M. E., Yang, X., Shadrach, K. G., Lu, L., Ufret, R. L., Salomon, R. G., and Perez, V. L. (2008). Oxidative damage–induced inflammation initiates age-related macular degeneration. *Nature Medicine*, 14(2):194–198.

Hosmer, D. W., Hosmer, T., Le Cessie, S., Lemeshow, S., et al. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.

Hosmer, D. W. and Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):1043–1069.

Hosmer Jr, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley-Interscience Publication, New York, NY, 2nd edition edition.

Hu, F. B. (2002). Dietary pattern analysis: a new direction in nutritional epidemiology. *Current Opinion in Lipidology*, 13(1):3–9.

Hu, F. B., Rimm, E., Smith-Warner, S. A., Feskanich, D., Stampfer, M. J., Ascherio, A., Sampson, L., and Willett, W. C. (1999). Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *The American Journal of Clinical Nutrition*, 69(2):243–249.

Hu, F. B., Rimm, E. B., Stampfer, M. J., Ascherio, A., Spiegelman, D., and Willett, W. C. (2000). Prospective study of major dietary patterns and risk of coronary heart disease in men. *The American Journal of Clinical Nutrition*, 72(4):912–921.

Hyman, L., Schachat, A. P., He, Q., and Leske, M. C. (2000). Hypertension, cardiovascular disease, and age-related macular degeneration. Age-Related Macular Degeneration Risk Factors Study Group. *Archives of Ophthalmology*, 118(3):351–358.

INSA (2006). Tabela da composição de alimentos. Lisboa: Instituto Nacional de Saúde Dr̲o Ricardo Jorge.

Islam, F. M. A., Chong, E. W., Hodge, A. M., Guymer, R. H., Aung, K. Z., Makeyeva, G. A., Baird, P. N., Hopper, J. L., English, D. R., Giles, G. G., et al. (2014). Dietary Patterns and Their Associations with Age-Related Macular Degeneration: The Melbourne Collaborative Cohort Study. *Ophthalmology*, 121(7):1428–1434.

Jager, R. D., Mieler, W. F., and Miller, J. W. (2008). Age-Related Macular Degeneration. *New England Journal of Medicine*, 358(24):2606–2617.

Janssen, I., Katzmarzyk, P. T., and Ross, R. (2002). Body mass index, waist circumference, and health risk: evidence in support of current National Institutes of Health guidelines. *Archives of Internal Medicine*, 162(18):2074–2079.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied Multivariate Statistical Analysis*, volume 5. Prentice Hall, Upper Saddle River, NJ, 5th edition.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, New York, 2nd edition.

Kahn, H. A., Leibowitz, H. M., Ganley, J. P., Kini, M. M., Colton, T., Nickerson, R. S., and Dawber, T. R. (1977). The Framingham Eye Study II. Association of ophthalmic pathology with single variables previously measured in the Framingham Heart Study. *American Journal of Epidemiology*, 106(1):33–41.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12(1):1–16.

Kant, A., Schatzkin, A., Block, G., Ziegler, R., and Nestle, M. (1991). Food group intake patterns and associated nutrient profiles of the US population. *Journal of the American Dietetic Association*, 91(12):1532–1537.

Kant, A. K. (1996). Indexes of overall diet quality: a review. *Journal of the American Dietetic Association*, 96(8):785–791.

Kant, A. K. (2004). Dietary patterns and health outcomes. *Journal of the American Dietetic Association*, 104(4):615–635.

Karesvuo, P., Gursoy, U. K., Pussinen, P. J., Suominen, A. L., Huumonen, S., Vesti, E., and Könönen, E. (2013). Alveolar bone loss associated with age-related macular degeneration in males. *Journal of Periodontology*, 84(1):58–67.

Khotcharrat, R., Patikulsila, D., Hanutsaha, P., Khiaocham, U., Ratanapakorn, T., Sutheerawatananonda, M., and Pannarunothai, S. (2015). Epidemiology of Age-Related Macular Degeneration among the Elderly Population in Thailand. *Journal of the Medical Association of Thailand*, 98(8):790–797.

Kiernan, D. F., Hariprasad, S. M., Rusu, I. M., Mehta, S. V., Mieler, W. F., and Jager, R. D. (2010). Epidemiology of the association between anticoagulants and intraocular hemorrhage in patients with neovascular age-related macular degeneration. *Retina*, 30(10):1573–1578.

Kim, G. H., Kim, J. E., Rhie, S. J., and Yoon, S. (2015). The Role of Oxidative Stress in Neurodegenerative Diseases. *Experimental Neurobiology*, 24(4):325–340.

Klein, B. E. and Klein, R. (1982). Cataracts and macular degeneration in older Americans. *Archives of Ophthalmology*, 100(4):571–573.

Klein, B. E., Klein, R., Lee, K. E., and Jensen, S. C. (2001). Measures of obesity and age-related eye diseases. *Ophthalmic Epidemiology*, 8(4):251–262.

Klein, R. (2007). Overview of progress in the epidemiology of age-related macular degeneration. *Ophthalmic Epidemiology*, 14(4):184–187.

Klein, R., Klein, B. E., and Jensen, S. C. (1997). The relation of cardiovascular disease and its risk factors to the 5-year incidence of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology*, 104(11):1804–1812.

Klein, R., Klein, B. E., Jensen, S. C., Mares-Perlman, J. A., Cruickshanks, K. J., and Palta, M. (1999). Age-related maculopathy in a multiracial United States population: the National Health and Nutrition Examination Survey III. *Ophthalmology*, 106(6):1056–1065.

Klein, R., Klein, B. E., and Linton, K. L. (1992). Prevalence of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology*, 99(6):933–943.

Klein, R., Klein, B. E., and Moss, S. E. (1998). Relation of Smoking to the Incidence of Age-related Maculopathy The Beaver Dam Eye Study. *American Journal of Epidemiology*, 147(2):103–110.

Klein, R., Peto, T., Bird, A., and Vannewkirk, M. R. (2004). The epidemiology of age-related macular degeneration. *American Journal of Ophthalmology*, 137(3):486–495.

Kleinbaum, D., Kupper, L., Nizam, A., and Rosenberg, E. (2013). *Applied regression analysis and other multivariable methods*. Cengage Learning, Boston, MA, 5th edition.

Knudtson, M. D., Klein, R., and Klein, B. E. (2006). Physical activity and the 15-year cumulative incidence of age-related macular degeneration: the beaver dam eye study. *British journal of ophthalmology*, 90(12):1461–1463.

Knudtson, M. D., Klein, R., and Klein, B. E. (2007). Alcohol consumption and the 15-year cumulative incidence of age-related macular degeneration. *American journal of ophthalmology*, 143(6):1026–1029.

Landau, S. and Ster, I. C. (2010). Cluster analysis: overview.

Lim, L. S., Mitchell, P., Seddon, J. M., Holz, F. G., and Wong, T. Y. (2012). Age-related macular degeneration. *The Lancet*, 379(9827):1728–1738.

Lopes, C. (2000). Reprodutibilidade e validação do questionário semiquantitativo de frequência alimentar. *Alimentação e Enfarte agudo do miocárdio: Estudo de caso-controlo de base comunitária*, pages 78–115.

Lopes, C., Oliveira, A., Santos, A. C., Ramos, E., Gaio, A. R., Severo, M., and Barros, H. (2006). Consumo alimentar no porto.

Lovric, M. (2011). *International Encyclopedia of Statistical Science*. Springer, Berlin, 2011 edition edition.

Magalhães, B., Peleteiro, B., and Lunet, N. (2012). Dietary patterns and colorectal cancer: systematic review and meta-analysis. *European journal of cancer prevention*, 21(1):15–23.

Maldonado, G. and Greenland, S. (1993). Simulation study of confounder-selection strategies. *American journal of epidemiology*, 138(11):923–936.

Maltzman, B. A., Mulvihill, M. N., and Greenbaum, A. (1979). Senile macular degeneration and risk factors: a case-control study. *Annals of ophthalmology*, 11(8):1197–1201.

Mares, J. A., Voland, R. P., Sondel, S. A., Millen, A. E., LaRowe, T., Moeller, S. M., Klein, M. L., Blodi, B. A., Chappell, R. J., Tinker, L., et al. (2011). Healthy lifestyles related to subsequent prevalence of age-related macular degeneration. *Archives of ophthalmology*, 129(4):470–480.

Martínez-González, M. A., García-López, M., Bes-Rastrollo, M., Toledo, E., Martínez-Lapiscina, E. H., Delgado-Rodriguez, M., Vazquez, Z., Benito, S., and Beunza, J. J. (2011). Mediterranean diet and the incidence of cardiovascular disease: a spanish cohort. *Nutrition, Metabolism and Cardiovascular Diseases*, 21(4):237–244.

McCance, R. (2002). Mccance & widdowson's the composition of foods: summary edition.

McEvoy, C. T., Cardwell, C. R., Chakravarthy, U., Hogg, R. E., McKinley, M. C., Young, I. S., Fletcher, A. E., and Woodside, J. V. (2013). A posteriori–derived dietary patterns and retinal vessel caliber in an elderly populationa posteriori–derived dietary patterns. *Investigative ophthalmology & visual science*, 54(2):1337–1344.

McGwin, G., Owsley, C., Curcio, C., and Crain, R. (2003). The association between statin use and age related maculopathy. *British journal of ophthalmology*, 87(9):1121–1125.

Merle, B., Delyfer, M.-N., Korobelnik, J.-F., Rougier, M.-B., Colin, J., Malet, F., Féart, C., Le Goff, M., Dartigues, J.-F., Barberger-Gateau, P., et al. (2011). Dietary omega-3 fatty acids and the risk for age-related maculopathy: the Alienor Study. *Investigative Ophthalmology & Visual Science*, 52(8):6004–6011.

Merle, B. M., Silver, R. E., Rosner, B., and Seddon, J. M. (2015). Adherence to a Mediterranean diet, genetic susceptibility, and progression to advanced macular degeneration: a prospective cohort study. *The American Journal of Clinical Nutrition*, 102(5):1196–1206.

Meyers, K. J., Liu, Z., Millen, A. E., Iyengar, S. K., Blodi, B. A., Johnson, E., Snodderly, D. M., Klein, M. L., Gehrs, K. M., Tinker, L., et al. (2015). Joint Associations of Diet, Lifestyle, and Genes with Age-Related Macular Degeneration. *Ophthalmology*, 122(11):2286–2294.

Millen, A. E., Meyers, K. J., Liu, Z., Engelman, C. D., Wallace, R. B., LeBlanc, E. S., Tinker, L. F., Iyengar, S. K., Robinson, J. G., Sarto, G. E., et al. (2015). Association between vitamin D status and age-related macular degeneration by genetic risk. *JAMA Ophthalmology*, 133(10):1171–1179.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Mitchell, P., Smith, W., Attebo, K., and Wang, J. J. (1995). Prevalence of age-related maculopathy in Australia: the Blue Mountains Eye Study. *Ophthalmology*, 102(10):1450–1460.

Moeller, S. M., Reedy, J., Millen, A. E., Dixon, L. B., Newby, P., Tucker, K. L., Krebs-Smith, S. M., and Guenther, P. M. (2007). Dietary patterns: challenges and opportunities in dietary patterns research: an Experimental Biology workshop, April 1, 2006. *Journal of the American Dietetic Association*, 107(7):1233–1239.

Mooi, E. and Sarstedt, M. (2010). Cluster analysis. In *A Concise Guide to Market Research*, pages 237–284. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-12541-6_9.

Moss, S. E., Klein, R., Klein, B. E., Jensen, S. C., and Meuer, S. M. (1998). Alcohol consumption and the 5-year incidence of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology*, 105(5):789–794.

Nelder, J. A. and Baker, R. J. (1972). Generalised Linear Model. *Journal of the Royal Statistical Society*, 135(3):370–384.

Newby, P., Muller, D., and Tucker, K. L. (2004). Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods. *The American Journal of Clinical Nutrition*, 80(3):759–767.

Newby, P. and Tucker, K. L. (2004). Empirically derived eating patterns using factor or cluster analysis: a review. *Nutrition Reviews*, 62(5):177–203.

Nitsch, D., Douglas, I., Smeeth, L., and Fletcher, A. (2008). Age-related Macular Degeneration and Complement Activation–related Diseases: A Population-Based Case-Control Study. *Ophthalmology*, 115(11):1904–1910.

Nonyane, B. A., Nitsch, D., Whittaker, J. C., Sofat, R., Smeeth, L., Chakravarthy, U., and Fletcher, A. E. (2010). An ecological correlation study of late age-related macular degeneration and the complement factor H Y402H polymorphism. *Investigative Ophthalmology & Visual Science*, 51(5):2393–2402.

Northstone, K., Ness, A., Emmett, P., and Rogers, I. (2008). Adjusting for energy intake in dietary pattern investigations using principal components analysis. *European Journal of Clinical Nutrition*, 62(7):931–938.

Obisesan, T. O., Hirsch, R., Kosoko, O., Carlson, L., and Parrott, M. (1998). Moderate Wine Consumption Is Associated with Decreased Odds of Developing Age-Related Macular Degeneration in NHANES-1. *Journal of the American Geriatrics Society*, 46(1):1–7.

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41(5):673–690.

Ocké, M. C. (2013). Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis. *Proceedings of the Nutrition Society*, 72(02):191–199.

Ottenbacher, K. J., Smith, P. M., Illig, S. B., Linn, R. T., Fiedler, R. C., and Granger, C. V. (2001). Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *Journal of Clinical Epidemiology*, 54(11):1159–1165.

Park, S. J., Lee, J. H., Woo, S. J., Ahn, J., Shin, J. P., Song, S. J., Kang, S. W., Park, K. H., of the Korean, E. S. C., and Society, O. (2014). Age-related macular degeneration: prevalence and risk factors from Korean National Health and Nutrition Examination Survey, 2008 through 2011. *Ophthalmology*, 121(9):1756–1765.

Paul, L. C. and Al Suman, A. (2013). Methodological Analysis of Principal Component Analysis Method. *International Journal of Scientific & Engineering Research*, 809:98.

Quintal, G. (2006). Análise de clusters aplicada ao Sucesso/Insucesso em Matemática. Master's thesis, Departamento de Matemática e Engenharias, Universidade da Madeira.

R Core Team and contributors worldwide (2017). The R Stats Package. R package version 3.5.0.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rai, P. (2011). Data Clustering: K-means and Hierarchical Clustering. School of Computing, The University of Utah.

Randall, E., Marshall, J. R., Brasure, J., and Graham, S. (1992). Dietary patterns and colon cancer in Western New York. *Nutrition and Cancer*, 18(3):265–276.

Reis, E. (2001). *Estatística Multivariada Aplicada*. Sílabo.

Reis dos Santos, P. M. and Reis dos Santos, M. I. (2012). Construction of stochastic simulation metamodels using smoothing splines. *International Journal of Simulation and Process Modelling*, 7(4):249–261.

Reynolds, K., Lewis, B., Nolen, J. D. L., Kinney, G. L., Sathya, B., and He, J. (2003). Alcohol consumption and risk of stroke: a meta-analysis. *JAMA*, 289(5):579–588.

Ritter, L. L., Klein, R., Klein, B. E., Mares-Perlman, J. A., and Jensen, S. C. (1995). Alcohol use and age-related maculopathy in the Beaver Dam Eye Study. *American Journal of Ophthalmology*, 120(2):190–196.

Robertson, C., Boyle, P., Hsieh, C.-c., Macfarlane, G. J., and Maisonneuve, P. (1994). Some statistical considerations in the analysis of case-control studies when the exposure variables are continuous measurements. *Epidemiology*, 5(2):164–170.

Rosenbloom, C. A. and Whittington, F. J. (1993). The effects of bereavement on eating behaviors and nutrient intakes in elderly widowed persons. *Journal of Gerontology*, 48(4):223–229.

Rosenfeld, P. J., Brown, D. M., Heier, J. S., Boyer, D. S., Kaiser, P. K., Chung, C. Y., and Kim, R. Y. (2006). Ranibizumab for neovascular age-related macular degeneration. *New England Journal of Medicine*, 355(14):1419–1431.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Number 12. Cambridge University Press, Cambridge, 1st edition.

Sarkar, S. K., Midi, H., Rana, M., et al. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences*, 11(1):26–35.

Schick, T., Ersoy, L., Lechanteur, Y. T., Saksens, N. T., Hoyng, C. B., Den Hollander, A. I., Kirchhof, B., and Fauser, S. (2016). History of Sunlight Exposure is a Risk Factor for Age-Related Macular Degeneration. *Retina*, 36(4):787–790.

Schottenfeld, D. and Fraumeni Jr, J. F. (2006). *Cancer Epidemiology and Prevention*. Oxford University Press, Oxford, 3rd edition.

Schramm, E. C., Clark, S. J., Triebwasser, M. P., Raychaudhuri, S., Seddon, J. M., and Atkinson, J. P. (2014). Genetic variants in the complement system predisposing to age-related macular degeneration: a review. *Molecular Immunology*, 61(2):118–125.

Schulze, M. B., Hoffmann, K., Kroke, A., and Boeing, H. (2003). An approach to construct simplified measures of dietary patterns from exploratory factor analysis. *British Journal of Nutrition*, 89(03):409–418.

Scotia, N. (2010). Explaining Odds Ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3):227.

Seddon, J. M., Cote, J., Davis, N., and Rosner, B. (2003a). Progression of age-related macular degeneration: association with body mass index, waist circumference, and waist-hip ratio. *Archives of Ophthalmology*, 121(6):785–792.

Seddon, J. M., Cote, J., and Rosner, B. (2003b). Progression of age-related macular degeneration: association with dietary fat, transunsaturated fat, nuts, and fish intake. *Archives of Ophthalmology*, 121(12):1728–1737.

Seddon, J. M., George, S., and Rosner, B. (2006). Cigarette smoking, fish consumption, omega-3 fatty acid intake, and associations with age-related macular degeneration: the US Twin Study of Age-Related Macular Degeneration. *Archives of Ophthalmology*, 124(7):995–1001.

Shalizi, C. (2009). Distances between Clustering, Hierarchical Clustering. Department of Statistics, Carnegie Mellon University.

Shaw, P. X., Stiles, T., Douglas, C., Ho, D., Fan, W., Du, H., and Xiao, X. (2016). Oxidative stress, innate immunity, and age-related macular degeneration. *AIMS Molecular Science*, 3(2):196.

Shim, J., Ryu, J., and Paik, H. (1997). Contribution of seasonings to nutrient intake assessed by food frequency questionnaire in adults in rural area of Korea. *Korean Journal of Community Nutrition*, 30(10):1211–1218.

Shmueli, G. et al. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Sin, H. P., Liu, D. T., and Lam, D. S. (2013). Lifestyle Modification, Nutritional and Vitamins Supplements for Age-Related Macular Degeneration. *Acta Ophthalmologica*, 91(1):6–11.

Siou, G. L., Yasui, Y., Csizmadi, I., McGregor, S. E., and Robson, P. J. (2011). Exploring statistical approaches to diminish subjectivity of cluster analysis to derive dietary patterns the tomorrow project. *American journal of epidemiology*, 173(8):956–967.

Slattery, M. L., Boucher, K. M., Caan, B. J., Potter, J. D., and Ma, K.-N. (1998). Eating patterns and risk of colon cancer. *American Journal of Epidemiology*, 148(1):4–16.

Smiderle, L. (2013). Atividade antioxidante, polifenóis totais, carotenoides totais, α- e β- carotenos e isômeros trans (e) e cis (z) em cultivares de abóbora (cucurbita moschata) cruas e cozidas.

Smith, W., Mitchell, P., Leeder, S. R., and Wang, J. J. (1998). Plasma fibrinogen levels, other cardiovascular risk factors, and age-related maculopathy: the Blue Mountains Eye Study. *Archives of Ophthalmology*, 116(5):583–587.

Sperduto, R. D. and Hiller, R. (1986). Systemic hypertension and age-related maculopathy in the Framingham Study. *Archives of Ophthalmology*, 104(2):216–219.

StataCorp (2013a). Stata 13 Base Reference Manual.

StataCorp (2013b). Stata Statistical Software: Release 13.

Sunness, J. S. (1999). The natural history of geographic atrophy, the advanced atrophic form of age-related macular degeneration. *Molecular Vision*, 5(5):25.

Swenor, B. K., Bressler, S., Caulfield, L., and West, S. K. (2010). The impact of fish and shellfish consumption on age-related macular degeneration. *Ophthalmology*, 117(12):2395–2401.

Tan, J. S., Wang, J. J., Flood, V., and Mitchell, P. (2009). Dietary fatty acids and the 10-year incidence of age-related macular degeneration: the Blue Mountains Eye Study. *Archives of Ophthalmology*, 127(5):656–665.

Thorpe, M. G., Milte, C. M., Crawford, D., and McNaughton, S. A. (2016). A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. *International Journal of Behavioral Nutrition and Physical Activity*, 13(1):1.

Tomany, S. C., Wang, J. J., van Leeuwen, R., Klein, R., Mitchell, P., Vingerling, J. R., Klein, B. E., Smith, W., and de Jong, P. T. (2004). Risk factors for incident age-related macular degeneration: pooled findings from 3 continents. *Ophthalmology*, 111(7):1280–1287.

Topouzis, F., Anastasopoulos, E., Augood, C., Bentham, G. C., Chakravarthy, U., De Jong, P. T., Rahu, M., Seland, J., Soubrane, G., Tomazzoli, L., et al. (2009). Association of diabetes with age-related macular degeneration in the EUREYE study. *British Journal of Ophthalmology*, 93(8):1037–1041.

Trichopoulos, D. and Lagiou, P. (2001). Dietary patterns and mortality. *British Journal of Nutrition*, 85(2):133–134.

Trichopoulou, A., Kouris-Blazos, A., Wahlqvist, M. L., Gnardellis, C., Lagiou, P., Polychronopoulos, E., Vassilakou, T., Lipworth, L., and Trichopoulos, D. (1995). Diet and overall survival in elderly people. *BMJ*, 311(7018):1457–1460.

Vaičaitienė, R., Lukšienė, D. K., Paunksnis, A., Černiauskienė, L. R., Domarkienė, S., and Cimbalas, A. (2003). Age-related maculopathy and consumption of fresh vegetables and fruits in urban elderly. *Medicina (Kaunas)*, 39:1231–6.

Varma, R., Fraser-Bell, S., Tan, S., Klein, R., Azen, S. P., Group, L. A. L. E. S., et al. (2004). Prevalence of age-related macular degeneration in Latinos: the Los Angeles Latino eye study. *Ophthalmology*, 111(7):1288–1297.

Velie, E. M., Schairer, C., Flood, A., He, J.-P., Khattree, R., and Schatzkin, A. (2005). Empirically derived dietary patterns and risk of postmenopausal breast cancer in a large prospective cohort study. *The American Journal of Clinical Nutrition*, 82(6):1308–1319.

Vingerling, J. R., Dielemans, I., Hofman, A., Grobbee, D. E., Hijmering, M., Kramer, C. F., and de Jong, P. T. (1995a). The prevalence of age-related maculopathy in the Rotterdam Study. *Ophthalmology*, 102(2):205–210.

Vingerling, J. R., Dielemans, I., Witteman, J. C., Hofman, A., Grobbee, D. E., and De Jong, P. T. (1995b). Macular degeneration and early menopause: a case-control study. *BMJ*, 310(6994):1570–1571.

Wang, S., Xu, L., Jonas, J. B., Wang, Y. X., You, Q. S., and Yang, H. (2012). Dyslipidemia and eye diseases in the adult Chinese population: the Beijing Eye Study. *PLoS One*, 7(3).

Whiting, M. G. and Leverton, R. M. (1960). Reliability of dietary appraisal: comparisons between laboratory analysis and calculation from tables of food values. *American Journal of Public Health and the Nations Health*, 50(6 Pt 1):815–823.

Wickham, H. and Chang, W. (2016). Create elegant data visualisations using the grammar of graphics. R package version 2.2.1.

Willett, W. (1998). Food Frequency Methods. *Nutritional Epidemiology*, 5.

Willett, W. C., Howe, G. R., and Kushi, L. H. (1997). Adjustment for total energy intake in epidemiologic studies. *The American Journal of Clinical Nutrition*, 65(4):1220–1228.

Willett, W. C., Sacks, F., Trichopoulou, A., Drescher, G., Ferro-Luzzi, A., Helsing, E., and Trichopoulos, D. (1995). Mediterranean diet pyramid: a cultural model for healthy eating. *The American Journal of Clinical Nutrition*, 61(6):1402–1406.

Williams, P. T. (2009). Prospective study of incident age-related macular degeneration in relation to vigorous physical activity during a 7-year follow-up. *Investigative Ophthalmology & Visual Science*, 50(1):101–106.

Wirfält, E., Drake, I., and Wallström, P. (2013). What do review papers conclude about food and dietary patterns? *Food & Nutrition Research*, 57.

Wright, R. E. (1995). Logistic Regression. In Grimm, L. G. and Yarnold, P. R., editors, *Reading & Understanding Multivariate Statistics*, volume 1980, pages 217–244. American Psychological Association, Washington, DC, US.

Wu, S. (2010). *Goodness-of-fit tests for logistic regression*. The Florida State University.

Xu, L., Xie, X., Wang, Y., and Jonas, J. (2009). Ocular and systemic factors associated with diabetes mellitus in the adult population in rural and urban China. The Beijing Eye Study. *Eye*, 23(3):676–682.

Yim, O. and Ramdeen, K. T. (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *Quantitative Methods for Psychology*, 11(1):8–21.

You, Q. S., Xu, L., Yang, H., Li, Y. B., Wang, S., Da Wang, J., Zhang, J. S., Wang, Y. X., and Jonas, J. B. (2012). Five-year incidence of age-related macular degeneration: the Beijing Eye Study. *Ophthalmology*, 119(12):2519–2525.

Yu, S. S., Tang, X., Ho, Y.-S., Chang, R. C.-C., and Chiu, K. (2016). Links between the Brain and Retina: The Effects of Cigarette Smoking-Induced Age-Related Changes in Alzheimer's Disease and Macular Degeneration. *Frontiers in Neurology*, 7:119.

Yun, S. H., Choi, B.-Y., and Kim, M.-K. (2009). The effect of seasoning on the distribution of nutrient intakes by a food-frequency questionnaire in a rural area. *Korean Journal of Nutrition*, 42(3):246–255.

Yun, S. H., Shim, J.-S., Kweon, S., and Oh, K. (2013). Development of a food frequency questionnaire for the Korea National Health and Nutrition Examination Survey: data from the fourth Korea National Health and Nutrition Examination Survey (KNHANES IV). *Korean Journal of Nutrition*, 46(2):186–196.

Apêndice A - Questionário semiquantitativo de frequência alimentar

Neste anexo é apresentado o questionário aplicado para a recolha de dados (Inquérito - estilos de vida e hábitos alimentares).

Inquérito - estilos de vida e hábitos alimentares

Data do Inquérito: ____/____/_____ Iniciais: |__|__|__| Número de Participante:|__|__|__|__|__|

Parte A – Estilos de vida

1.Sexo: Feminino ☐ Masculino ☐

2.Data nascimento: ____/____/_____

3.Profissão: _____ 4.Escolaridade: _____

5.Peso: _____Kg Altura: _____ m Perímetro abdominal: _____ cm

6.Patologias Diagnosticadas:

	Sim	Não
Diabetes Mellitus	☐	☐
Hipertensão arterial	☐	☐
Dislipidémia (hipercolesterolemia, hipertrigliceridemia, etc)	☐	☐
Obesidade/Excesso de peso	☐	☐

Outra patologia. Qual?_____

7. Pratica algum tipo de Exercício Físico?

- Nenhum ☐
- Futebol ☐
- Basquetebol ☐
- Natação ☐
- Mais do que uma modalidade ☐
- Outras. Qual? _____

Quantas horas por semana? _____

8. Toma algum suplemento de vitaminas e/ou minerais ou outro suplemento alimentar?

Sim ☐ **Não** ☐ Se sim: Qual e qual o motivo?_____

Nº Comprimidos/dia ou dose diária: _____

9. Fuma? Sim ☐ **Nunca fumou** ☐ **Deixou de fumar** ☐

Número de cigarros /dia em média _____ Número de anos que fumou _____
(1 charuto=2 cigarros; 1 cachimbo=4 cigarros)

10. Normalmente quantas refeições faz por dia? _____

11. Toma o pequeno almoço diariamente? Sim ☐ **Não** ☐

12. No quotidiano, qual (quais) o (s) tipo (s) de alimentação predominante (s)?

- Refeição tradicional (sopa, prato, sobremesa) **Sim** ☐ **Não** ☐
- Fast food (hambúrgueres, pizzas, cachorros, etc) **Sim** ☐ **Não** ☐
- Alimentos pré confeccionados (rissóis, refeições congeladas, etc) **Sim** ☐ **Não** ☐
- Pratos combinados, baguetes **Sim** ☐ **Não** ☐
- Refeições ligeiras (saladas, sopas, etc) **Sim** ☐ **Não** ☐
- Outro tipo. Qual?_____

ID 0 0 0 | | | | | 0 0

Por favor, **antes de iniciar o questionário leia as instruções da página anterior.**
Pense durante o último ano quantas vezes por dia, semana ou mês, em média, consumiu cada um dos alimentos referidos. Na coluna referente à quantidade deverá assinalar se sua porção é igual, menor ou maior do que a referida como porção média. Para os alimentos consumidos só em determinadas épocas do ano, anote a frequência com que o alimento é consumido nessa época e assinale com uma cruz **(x)** na última coluna (Sazonal).

I. P. LÁCTEOS	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
1. Leite gordo	○	○	○	○	○	○	○	○	○	1 chávena = 250 ml	○	○	○	□
2. Leite meio-gordo	○	○	○	○	○	○	○	○	○	1 chávena = 250 ml	○	○	○	□
3. Leite magro	○	○	○	○	○	○	○	○	○	1 chávena = 250 ml	○	○	○	□
4. Iogurte	○	○	○	○	○	○	○	○	○	Um =125g	○	○	○	□
5.Queijo (de qualquer tipo incluindo queijo fresco e requeijão)	○	○	○	○	○	○	○	○	○	1 fatia = 30g	○	○	○	□
6. Sobremesas lácteas: pudim, aletria e leite creme , etc	○	○	○	○	○	○	○	○	○	Um ou 1 prato sobremesa	○	○	○	□
7. Gelados	○	○	○	○	○	○	○	○	○	Um ou 2 bolas	○	○	○	□

II. OVOS, CARNES E PEIXES	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
8.Ovos	○	○	○	○	○	○	○	○	○	Um	○	○	○	□
9.Frango	○	○	○	○	○	○	○	○	○	1 porção ou 2 peças=150g	○	○	○	□
10.Peru, coelho	○	○	○	○	○	○	○	○	○	1 porção ou 2 peças=150g	○	○	○	□
11.Carne vaca, porco,cabrito	○	○	○	○	○	○	○	○	○	1 porção =120g	○	○	○	□
12. Figado de vaca, porco, frango	○	○	○	○	○	○	○	○	○	1 porção = 120g	○	○	○	□
13 Língua, mão de vaca, tripas, chispe, coração, rim	○	○	○	○	○	○	○	○	○	1 porção =100g	○	○	○	□
14.Fiambre, chouriço, salpição, presunto, etc	○	○	○	○	○	○	○	○	○	2 fatias ou 3 rodelas =20g	○	○	○	□
15. Salsichas	○	○	○	○	○	○	○	○	○	3 médias	○	○	○	□
16. Toucinho, bacon	○	○	○	○	○	○	○	○	○	2 fatias=50g	○	○	○	□
17. Peixe gordo: sardinha, cavala, carapau, salmão,	○	○	○	○	○	○	○	○	○	1 porção =125g	○	○	○	□
18. Peixe magro: pescada, faneca, dourada, etc	○	○	○	○	○	○	○	○	○	1 porção =125g	○	○	○	□
19.Bacalhau	○	○	○	○	○	○	○	○	○	1 porção =125g	○	○	○	□
20.Peixe conserva: atum, sardinhas,etc	○	○	○	○	○	○	○	○	○	1 lata	○	○	○	□
21.Lulas, polvo	○	○	○	○	○	○	○	○	○	1 porção =100g	○	○	○	□
22.Camarão, amêijoas, mexilhão, etc	○	○	○	○	○	○	○	○	○	1 prato sobremesa =100g	○	○	○	□

III. Óleos e Gorduras	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
23. Azeite	○	○	○	○	○	○	○	○	○	1 colher sopa	○	○	○	□
24. Óleos: girassol, milho, soja	○	○	○	○	○	○	○	○	○	1 colher sopa	○	○	○	□
25. Margarina	○	○	○	○	○	○	○	○	○	1 colher chá	○	○	○	□
26. Manteiga	○	○	○	○	○	○	○	○	○	1 colher chá	○	○	○	□

Unidade de Epidemiologia Nutricional
Serviço de Higiene e Epidemiologia - FMUP

ID 0 0 0 | | | | | 0 0

IV.PÃO, CEREAIS E SIMILARES	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
27. Pão branco ou tostas	○	○	○	○	○	○	○	○	○	Um ou 2 tostas = 40g	○	○	○	☐
28. Pão (ou tostas), integral,centeio, mistura	○	○	○	○	○	○	○	○	○	Um ou 2 tostas= 50g	○	○	○	☐
29. Broa, broa de avintes	○	○	○	○	○	○	○	○	○	1 fatia = 80g	○	○	○	☐
30. Flocos cereais (muesli, corn-flakes, chocapic,etc.)	○	○	○	○	○	○	○	○	○	1 chávena =40g	○	○	○	☐
31. Arroz	○	○	○	○	○	○	○	○	○	½ prato = 100g	○	○	○	☐
32. Massas: esparguete, macarrão, etc.	○	○	○	○	○	○	○	○	○	½ prato = 100g	○	○	○	☐
33. Batatas fritas caseiras	○	○	○	○	○	○	○	○	○	½ prato = 100g	○	○	○	☐
34. Batatas fritas de pacote	○	○	○	○	○	○	○	○	○	1 pacote pequeno =30g	○	○	○	☐
35. Batatas cozidas, assadas, estufadas e puré	○	○	○	○	○	○	○	○	○	2 batatas médias =160 g	○	○	○	☐

V.DOCES E PASTÉIS	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
36. Bolachas tipo maria, água e sal ou integrais	○	○	○	○	○	○	○	○	○	3 bolachas	○	○	○	☐
37. Outras bolachas ou biscoitos	○	○	○	○	○	○	○	○	○	3 bolachas	○	○	○	☐
38. Croissant, pasteis, bolicao, doughnut ou bolos	○	○	○	○	○	○	○	○	○	Um; 1 fatia = 80g	○	○	○	☐
39.Chocolate (tablete ou em pó)	○	○	○	○	○	○	○	○	○	3 quadrados; 1 colher sopa	○	○	○	☐
40. Snacks de chocolate (Mars, Twix, Kit Kat, etc)	○	○	○	○	○	○	○	○	○	Um	○	○	○	☐
41. Marmelada, compota, geleia, mel	○	○	○	○	○	○	○	○	○	1 colher sobremesa	○	○	○	☐
42. Açúcar	○	○	○	○	○	○	○	○	○	1 colher sobremesa; 1 pacote	○	○	○	☐

VI. HORTALIÇAS E LEGUMES	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
43. Couve branca, couve lombarda	○	○	○	○	○	○	○	○	○	½ chávena =75g	○	○	○	☐
44. Penca, Tronchuda	○	○	○	○	○	○	○	○	○	½ chávena = 65g	○	○	○	☐
45. Couve galega	○	○	○	○	○	○	○	○	○	½ chávena =65g	○	○	○	☐
46. Brócolos	○	○	○	○	○	○	○	○	○	½ chávena =85g	○	○	○	☐
47. Couve-flor, Couve-bruxelas	○	○	○	○	○	○	○	○	○	½ chávena =65g	○	○	○	☐
48. Grelos, Nabiças, Espinafres	○	○	○	○	○	○	○	○	○	½ chávena =72g	○	○	○	☐
49. Feijão verde	○	○	○	○	○	○	○	○	○	½ chávena =65g	○	○	○	☐
50. Alface, Agrião	○	○	○	○	○	○	○	○	○	½ chávena =15g	○	○	○	☐
51. Cebola	○	○	○	○	○	○	○	○	○	½ média=40g	○	○	○	☐
52. Cenoura	○	○	○	○	○	○	○	○	○	1 média=80g	○	○	○	☐
53. Nabo	○	○	○	○	○	○	○	○	○	1 médio=78g	○	○	○	☐
54. Tomate fresco	○	○	○	○	○	○	○	○	○	½ médio=63g	○	○	○	☐
55. Pimento	○	○	○	○	○	○	○	○	○	½ médio=68g	○	○	○	☐
56. Pepino	○	○	○	○	○	○	○	○	○	¼ médio=50g	○	○	○	☐
57. Leguminosas: feijão, grão de bico	○	○	○	○	○	○	○	○	○	1 chávena	○	○	○	☐
58. Ervilha grão, Fava	○	○	○	○	○	○	○	○	○	½ chávena	○	○	○	☐

Unidade de Epidemiologia Nutricional
Serviço de Higiene e Epidemiologia - FMUP

ID 0 0 0 | | | | | 0 0

VII. FRUTOS	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
59. Maça, pêra	○	○	○	○	○	○	○	○	○	uma média	○	○	○	□
60. Laranja, Tangerinas	○	○	○	○	○	○	○	○	○	1 média; 2 médias	○	○	○	□
61. Banana	○	○	○	○	○	○	○	○	○	uma média	○	○	○	□
62. Kiwi	○	○	○	○	○	○	○	○	○	um médio	○	○	○	□
63. Morangos	○	○	○	○	○	○	○	○	○	1 chávena	○	○	○	□
64. Cerejas	○	○	○	○	○	○	○	○	○	1 chávena	○	○	○	□
65. Pêssego, Ameixa	○	○	○	○	○	○	○	○	○	1 médio; 3 médios	○	○	○	□
66. Melão, Melancia	○	○	○	○	○	○	○	○	○	1 fatia média = 150g	○	○	○	□
67. Diospiro	○	○	○	○	○	○	○	○	○	1 médio	○	○	○	□
68. Figo fresco, Nêsperas, Damascos	○	○	○	○	○	○	○	○	○	3 médios	○	○	○	□
69. Uvas frescas	○	○	○	○	○	○	○	○	○	1 cacho médio	○	○	○	□
70. Frutos conserva pêssego, ananás	○	○	○	○	○	○	○	○	○	2 metades ou rodelas	○	○	○	□
71. Amêndoas, avelãs, nozes, amendoins, pistachio, etc.	○	○	○	○	○	○	○	○	○	½ chávena (descascado)	○	○	○	□
72. Azeitonas	○	○	○	○	○	○	○	○	○	6 unidades	○	○	○	□

VIII. BEBIDAS E MISCELANEAS	Frequência alimentar									Quantidade				sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	A sua porção é: Menor	Igual	Maior	
73. Vinho	○	○	○	○	○	○	○	○	○	1 copo=125ml	○	○	○	□
74. Cerveja	○	○	○	○	○	○	○	○	○	1 garrafa ou 1 lata=330 ml	○	○	○	□
75. Bebidas brancas: whisky, aguardente, brandy, etc	○	○	○	○	○	○	○	○	○	1 cálice = 40 ml	○	○	○	□
76. Coca-cola, pepsi-cola ou outras colas	○	○	○	○	○	○	○	○	○	1 garrafa ou 1 lata=330 ml	○	○	○	□
77. Ice-tea	○	○	○	○	○	○	○	○	○	1 garrafa ou 1 lata=330 ml	○	○	○	□
78.Outros refrigerantes, sumos de fruta ou néctares embalados	○	○	○	○	○	○	○	○	○	1 garrafa ou 1 copo = 250 ml	○	○	○	□
79.Café (incluindo pingo, meia de leite e outras bebidas com café)	○	○	○	○	○	○	○	○	○	1 chávena café	○	○	○	□
80. Chá preto e verde	○	○	○	○	○	○	○	○	○	1 chávena	○	○	○	□
81. Croquetes, rissóis, bolinhos de bacalhau, etc.	○	○	○	○	○	○	○	○	○	3 unidades	○	○	○	□
82. Maionese	○	○	○	○	○	○	○	○	○	1 colher sobremesa	○	○	○	□
83. Molho de tomate, ketchup	○	○	○	○	○	○	○	○	○	1 colher sopa	○	○	○	□
84. Pizza	○	○	○	○	○	○	○	○	○	Meia pizza-normal	○	○	○	□
85. Hambúrguer	○	○	○	○	○	○	○	○	○	Um médio	○	○	○	□
86. Sopa de legumes	○	○	○	○	○	○	○	○	○	1 prato	○	○	○	□

Existe algum alimento ou bebida que eu não tenha mencionado e que tenha consumido pelo menos 1 vez por semana mesmo em pequenas quantidades, ou numa época em particular. Por ex: **frutos tropicais, sumos de fruta natural, bebidas espirituosas, café de mistura, alheiras, farinheiras, frutos secos (figo, ameixa, damasco), produtos dietéticos, rebuçados, etc.**

Outros Alimentos	Frequência alimentar									Quantidade	sazonal
	Nunca ou <1 mês	1-3 por mês	1 por sem	2-4 por sem	5-6 por sem	1 por dia	2-3 por dia	4-5 por dia	6 + por dia	Porção Média	
	○	○	○	○	○	○	○	○	○		□
	○	○	○	○	○	○	○	○	○		□
	○	○	○	○	○	○	○	○	○		□

Unidade de Epidemiologia Nutricional
Serviço de Higiene e Epidemiologia - FMUP

Apêndice B - Tabelas das variáveis

Neste anexo são apresentadas tabelas referentes à descrição dos itens alimentares, correspondentes produtos na Tabela da Composição de Alimentos e códigos associados (INSA, 2006); à descrição das variáveis associadas a esses itens alimentares e respetivos valores de porção média. São apresentadas também tabelas com a descrição dos grupos alimentares utilizados na análise estatística e respetivas variáveis; com a descrição segundo o INSA dos nutrientes e variáveis associadas; e com a descrição das variáveis referentes às características socioeconómicas, sociodemográficas e respeitantes à dieta e ao estilo de vida dos participantes.

Tabela 7.1: Descrição dos itens alimentares do inquérito e correspondentes código e descrição segundo a Tabela da Composição de Alimentos do INSA utilizados

Item alimentar	**Código**	**Descrição INSA**
Leite gordo	IS023	Leite Vaca UHT gordo
Leite meio-gordo	IS025	Leite Vaca UHT meio gordo
Leite magro	IS027	Leite Vaca UHT magro
Iogurte	IS074	Iogurte Aromatizado açucarado sólido meio gordo
Queijo (de qualquer tipo, incluindo queijo fresco e requeijão)	IS050	Queijo Flamengo 45% gordura
Sobremesas láteas: pudim, aletria, leite creme	IS491	Pudim flan caseiro
Gelado	IS514	Gelado de leite
Ovos	IS086	Ovo (de galinha) cozido
Frango	IS244	Frango Inteiro com pele grelhado
Perú, coelho	IS286	Peru Perna com pele estufada com margarina
Carnes de vaca, porco, cabrito	IS220	Vaca Bife (valor médio de acém, alcatra e lombo) frito, sem molho
Fígado de vaca, porco, frango	IS337	Fígado de porco, frito, sem molho
Língua, mão de vaca, tripas, chispe, coração, rim	IS331	Língua de vaca, estufada, sem molho
Fiambre, chouriço, salpicão, presunto	IS358	Fiambre
Salsichas	IS364	Salsicha tipo Frankfurt
Toucinho, bacon	IS309	Porco Toucinho entremeado ligeiramente salgado, cozido sem adição de sal
Peixes gordos: sardinha, cavala, carapau, salmão	IS819	Carapau frito
Peixes magros: pescada, faneca, dourada	IS846	Linguado grelhado
Bacalhau	IS805	Bacalhau Seco e salgado, demolhado cozido
Peixe em conserva: atum, sardinhas	IS814	Atum conserva em óleo
Lulas, polvo	IS915	Lula grelhada
Camarão, amêijoas, mexilhão	IS910	Mexilhão cozido sem sal
Azeite	IS395	Azeite (4 marcas)
Oleos: girassol, milho, soja	IS390	Óleo de girassol
Margarina	IS380	Margarina culinária para folhados, com sal
Manteiga	IS385	Manteiga com sal
Pão branco ou tostas	IS429	Pão de trigo

Continua na próxima página

Tabela 7.1 Continuação

Item alimentar	Código	Descrição INSA
Pão ou tostas integrais, pão de centeio, pão de mistura	IS433	Pão de trigo integral
Broa, broa de Avintes	IS428	Pão de milho
Flocos de cereais (muesli, corn-flakes, chocapic)	IS443	Flocos de milho tipo "Corn Flakes"
Arroz	IS403	Arroz cozido simples
Massas: esparguete, macarrão	IS419	Esparguete cozido
Batatas fritas caseiras	IS591	Batata frita caseira (em palitos)
Batatas fritas de pacote	IS592	Batata frita, de pacote (em rodelas)
Batatas cozidas, assadas, estufadas, puré	IS586	Batata cozida
Bolachas maria, bolachas água e sal, bolachas integrais	IS461	Bolacha água e sal
Outras bolachas, biscoitos	IS465	Bolacha chocolate
Croissants, pastéis, Bolicao, donut, bolos	IS480	Croissant
Chocolates (tablete ou em pó)	IS507	Chocolate em pó
Snacks de chocolate (Mars, Twix, Kit-Kat)	IS506	Chocolate em barra, culinária
Marmelada, geleia, compota, mel	IS672	Marmelada
Açúcar	IS503	Açúcar branco
Couve branca, couve lombarda	IS561	Couve lombarda cozida
Couve penca, couve tronchuda	IS563	Couve portuguesa cozida
Couve galega	IS559	Couve galega cozida
Brocolos	IS551	Brócolos cozidos
Couve-flor, couve-bruxelas	IS557	Couve-flor cozida
Grelos, nabiças, espinafres	IS566	Grelos de couve cozidos
Feijão verde	IS578	Feijão verde fresco cozido
Alface, agrião	IS584	Alface crua
Cebola	IS598	Cebola cozida
Cenoura	IS601	Cenoura cozida
Nabo	IS610	Nabo (raiz) cozido
Tomate fresco	IS615	Tomate cru
Pimento	IS613	Pimento grelhado
Pepino	IS611	Pepino cru
Leguminosas secas: feijão, grão de bico	IS536	Grão-de-bico cozido (demolhado)
Ervilhas grão, favas	IS528	Favas secas cozidas (demolhadas)
Maçã, pera	IS662	Maçã com casca
Laranja, tangerina	IS691	Tangerina
Banana	IS636	Banana
Kiwi	IS657	Kiwi
Morangos	IS676	Morango
Cerejas	IS637	Cereja (4 variedades)
Pêssego, ameixa	IS685	Pêssego (2 variedades)
Melão, melancia	IS674	Melão (3 variedades)
Diospiro	IS649	Dióspiro
Figo fresco, nêspera, damasco	IS650	Figo (5 variedades)
Uvas frescas	IS694	Uva tinta (5 variedades)
Fruta em conserva (pêssego, ananás)	IS633	Ananás, conserva em calda de açúcar
Amêndoas, avelãs, amendoins, nozes, pistachios	IS697	Amêndoa, miolo, com pele
Azeitonas	IS703	Azeitona
Vinho	IS714	Alcoólicas Fermentadas - Vinho maduro tinto
Cerveja	IS726	Alcoólicas Fermentadas - Cerveja branca
Bebidas brancas: whisky, aguardente, brandy, vinho do Porto, licores	IS729	Alcoólicas Destiladas - Aguardente
Coca-Cola, Pepsi, outras colas	IS763	Não Alcoólicas, Bebida Refrigerante cola
Ice-tea	IS744	Não Alcoólicas, Sumo fresco de limão (espremido)
Outros refrigerantes, sumos de fruta, néctares embalados	IS750	Não Alcoólicas, Néctar laranja
Café (incluindo pingo, meia de leite e outras bebidas com café)	IS771	Não Alcoólicas, Café solúvel (pó) com cafeína (2 marcas)
Chá preto, chá verde	IS962	Não Alcoólicas, Chá, infusão, preto
Croquetes, rissóis, bolinhos de bacalhau	IS368	Rissol
Maionese	IS926	Maionese caseira, com ovo e azeite
Ketchup, molho de tomate	IS960	Molho de tomate, "Ketchup"
Pizza	IS955	Pizza de queijo, tomate e fiambre
Hambúrguer	IS294	Hamburger de vaca, grelhado
Sopa de legumes	IS792	Sopa feijão verde

Tabela 7.2: Tabela da Composição de Alimentos referente aos itens alimentares do inquérito. Valores por 100g parte edível (excepto para as bebidas alcoólicas; valores por 100ml parte edível)

Item alimentar / Grupo de alimentos	Energia (kcal)	Total HC disponíveis	Mono + dissacarídeos (g)	Oligossacarídeos (g)	Amido (g)	Fibra alimentar (g)	Gordura total (g)	Ácidos gordos monoinsaturados (g)	Ácidos gordos polinsaturados (g)	Ácido linoleico (g)	Ácidos gordos *trans* (g)	Ácidos gordos saturados (g)	Colesterol (mg)	Retinol (Vit.A total) (mg)	Vitamina C (mg)	α-tocoferol (Vit. E) (mg)	Caroteno (mg)	Cálcio (mg)	Ferro (mg)	Magnésio (mg)	Zinco (mg)	Proteína (g)
Iogurte	70.97498338	10.1	10.1	0	0	0	1.6	0.4	0.05	0.04	0.06	0.9	6	38	0	0.03	22	130	0.2	12	0.5	4.1
Queijo curado. semi-curado ou cremoso	316.2066945	0.2	0.2	0	0	0	23.4	6	0.9	0.7	1.06	12.6	69	268	0	0.44	201	800	0.8	40	5.3	26
Bacalhau	105.6260695	0	0	0	0	0	0.1	0	0.1	0	0	0	72	3	0	0.28	0	46	0.6	31	1.1	26.2
Azeite	900	0	0	0	0	0	99.9	78.6	6.9	6.2	0	14.4	0	0	0	14	0	0	0	0	0	0.1
Margarina	771.4048698	0.4	0.4	0	0	0	85.5	26.1	13.2	12.4	3.55	31.4	225	157	0	36	0	3	0.3	1	0.1	0.1
Manteiga	739.1435137	0.7	0.7	0	0	0	81.8	18.9	2.4	2	3.28	46.3	230	565	0	2	45	15	0.2	2	0.1	0.1
Açúcar	390.9598411	99.3	99.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
Vinho	65.47860419	0.2	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	1	9	0.1	0.1
Café (incluindo o adicionado a outras bebidas)	224.6346275	41.1	24.3	0	16.8	2	0.5	0	0.2	0.2	0	0.2	0	0	0	0	0	89	0.8	219	0.2	14.9
Sopa de legumes	39.19157331	5.4	0.8	0.04	4.6	0.8	1.5	1.1	0.1	0.1	0	0.2	0	14	4	0.3	86	11	0.2	8	0.1	0.9
Leite																						
Leite gordo	61.894	4.7	4.7	0	0	0	3.5	0.8	0.1	0.1	0.12	2	13	59	0	0.07	29	109	0.1	9	0.4	3
Leite meio-gordo	46.839	4.9	4.9	0	0	0	1.6	0.4	0.04	0.04	0.053	0.9	8	22	0	0.03	12	112	0.1	9	0.5	3.3
Leite magro	34	4.9	4.9	0	0	0	0.2	0.1	0	0	0	0.1	1	0	0	0	0	114	0.1	10	0.4	3.4
Carnes brancas																						
Frango	239.1	0	0	0	0	0	14.3	4.7	2.9	2.6	0.07	3.4	138.5	20	0	0.3	0	15	1.1	25	1.3	27.6
Perú. coelho	169.9098087	1.4	1.2	0.2	0	0.6	9.8	3.7	2.2	1.7	0.058	3.4	85	20	6	0.7	107	26	1.9	28	3	19.1
Carnes vermelhas																						
Carne vaca. porco. cabrito como prato principal	182.6749821	0	0	0	0	0	7.5	2.9	0.3	0.1	0.35	3.5	89	0	0	0.03	0	13	1.9	29	5	28.8
Fígado de vaca. porco. frango	179.9753518	0	0	0	0	0	8.8	3.1	1.4	0.9	0.024	3.1	256	10280	27	0.4	0	18	9.4	37	3.6	25.2
Língua.mão de vaca. tripas. chispe. coração rim	267.1718229	0	0	0	0	0	20.9	9.3	0.7	0.7	0.93	8	95	0	3	0.4	0	9	6.2	24	3.9	19.8
Snacks																						
Fiambre. chouriço. salpicão. presunto. etc	303.2567472	0.5	0.5	0	0	0	25.5	11.8	3	2.7	0.07	8.9	64	0	0	0.16	0	18	0.9	25	3	18
Salsichas e similares	177.5569449	2.4	1.2	0	1.2	0.1	14.7	5.7	1.6	1.4	0.03	4.8	46	0	0	0.3	0	14	1	4	0.5	9
Toucinho. bacon	371.849074	0	0	0	0	0	33.9	13.1	5.2	4.5	0.203	11.4	71	0	0	0	0	17	0.4	14	1.9	16.7
Ovos (Um)	149.119157	0	0	0	0	0	10.8	3.9	2.1	1.9	0.02	2.7	408	170	0	2.3	0	44	2.1	11	1.3	13
Pizza	224.1566815	23.5	1.6	0	22	1.6	10	4.6	0.9	0.8	0.214	3.3	17	50	4	1	166	175	1	26	1.2	9.2
Hambúrguer	183.0533241	0	0	0	0	0	8.2	3.7	0.3	0.3	0.37	3.2	86	0	0	0.1	0	13	2.2	29	4.9	27.3
Croquetes. rissóis. bolinhos de bacalhau. etc.	280.5543114	31.8	1.1	0	30.7	1.3	13.4	3.9	4.1	4.1	0.18	3.4	66	29	0	2.4	1	31	1	21	1.2	7.3
Peixe (gordo e magro)																						
Peixe gordo: sardinha. cavala. carapau. etc	186.3989462	1	0	0	1	0	9.6	2.5	4.2	3	0	1.5	39	8	0	0.79	0	86	2	37	1.1	24
Peixe magro: pescada. faneca. linguado. etc	94.15536515	0	0	0	0	0	0.2	0	0.1	0	0	0	51	5	0	0.39	0	25	0.3	36	0.7	23.1
Moluscos, crustáceos e peixe de conserva																						
Peixe conserva: atum. sardinhas. etc	214.1198152	0	0	0	0	0	13	3.8	7.1	6.8	0.177	0.9	41	23	0	1.9	0	9	0.7	40	0.9	24.3
Lulas. polvo	144.3396968	0	0	0	0	0	1.6	0.1	0.6	0.04	0	0.5	260	17	0	1.8	0	28	1	49	0.7	32.5
Camarão. amêijoas. mexilhão. etc.	96.54509523	2.8	0	0	2.8	0	2.1	0.3	0.8	0.02	0	0.4	56	360	0	1	0	78	4	42	4.1	16.8

Continua na próxima página

Tabela 7.2 Continuação

Item alimentar / Grupo de alimentos	Energia (kcal)	Total HC disponíveis	Mono + dissacarídeos (g)	Oligossacarídeos (g)	Amido (g)	Fibra alimentar (g)	Gordura total (g)	Ácidos gordos monoinsaturados (g)	Ácidos gordos polinsaturados (g)	Ácido linoleico (g)	Ácidos gordos *trans* (g)	Ácidos gordos saturados (g)	Colesterol (mg)	Retinol (Vit.A total) (mg)	Vitamina C (mg)	α-tocoferol (Vit. E) (mg)	Caroteno (mg)	Cálcio (mg)	Ferro (mg)	Magnésio (mg)	Zinco (mg)	Proteína (g)
Pão branco, integral, tostas, broa																						
Pão branco ou tostas	288.9183667	57.3	2.1	0	55.2	3.8	2.2	0.3	0.8	0.8	0	0.5	0	0	0	0	0	43	2.2	31	1	8.4
Pão integral ou tostas integrais	221.275	39.9	2.2	0	37.7	7.4	3	0.4	1.1	1.1	0	0.7	0	0	0	0.2	0	55	3	93	2	7.6
Broa. broa de avintes	185.4430542	37.2	0	0	37.2	3.7	1.2	0.3	0.6	0.6	0	0.2	0	0	0	0	0	14	1.3	37	0.4	5.3
Cereais, bolachas integrais																						
Flocos de cereais	374.4707035	81.1	6.2	0	74.9	3.9	1.1	0.4	0.3	0.27	0.14	0.3	0	0	0	0.4	0	2	1	14	0.3	7.9
Bolachas tipo maria ou água e sal	450.7030931	61	1.4	0	59.6	3.2	17.8	5.8	2.7	2.6	0.16	7.6	0	0	0	1.3	0	27	1.4	19	0.6	9.8
Arroz, massa, batatas cozidas, assadas																						
Arroz cozinhado	127.3726133	28	0	0	28	0.8	0.2	0.1	0.1	0	0	0	0	0	0	0	0	7	0.2	15	0.6	2.5
Massas: esparguete. macarrão cozinhadas	100.8466094	19.9	0.9	0	19	1.5	0.6	0.1	0.3	0.3	0	0.1	0	0	0	0	0	9	0.5	7	0.3	3.4
Batatas cozidas. puré. assadas	85.31336386	18.5	1.2	0	17.3	1.6	0	0	0	0	0	0	0	0	11	0.06	0	9	0.2	13	0.2	2.4
Batatas fritas																						
Batatas fritas caseiras	225.1125735	27.6	1.7	0	25.9	2.4	10.8	2.4	5.1	5	0.06	1.4	0	0	13	3.8	0	14	0.3	20	0.3	3.7
Batatas fritas de pacote	526.2185636	39	0.6	0	38.4	10.7	38.1	13.5	6.9	6.4	0.36	14.7	0	0	27	5.5	0	21	1.6	45	1.7	5.7
Doces																						
Outras bolachas ou biscoitos	465.7583926	65.4	24	0	41.4	3	19.8	5.6	1.3	1.3	0.17	12.4	0	0	0	1.1	0	16	1.2	41	0.8	5.7
Croissant. pastéis ou bolos caseiros	415.8130339	42.2	0.6	0	41.6	2.6	23.5	7	2.5	2.5	1.75	10.4	52	21	0	0.06	0	45	1.8	30	0.8	7.6
Chocolate (tablete ou em pó)	452	63.8	60.5	0	3.3	7.3	20.3	6.8	0.7	0.7	0	12.8	0	7	0	0.18	39	42	2.2	104	1.1	4.2
Snacks de chocolate (Mars. Twix. Kit Kat. etc)	469.3429877	44	44	0	0	15	30.5	10.2	1	1	0	19.3	0	6	0	0.5	38	43	2.9	106	2	5.4
Marmelada compota.geleia. mel	270.9953911	69.7	69.7	0	0	2.2	0	0	0	0	0	0	0	0	0	0	0	8	2	5	0.1	0.1
Sobremesas lácteas: pudim flan. pudim de chocolate. etc	207.667544	36.4	36.4	0	0	0	4.5	1.5	0.7	0.7	0.03	1.4	142	103	0	0.7	6	76	1.1	10	0.7	6
Gelados	197.625	21.7	21.7	2	0	0	10.9	2.5	0.3	0.3	0.37	6.1	33	115	1	0.2	195	140	0.1	14	0.4	3.6
Hortícolas																						
Couve branca. C.lombarda cozinhadas	15.77221853	1.4	1.3	0	0.1	2.9	0.2	0	0.1	0.1	0	0	0	104	44	0.2	625	46	0.4	8	0.2	2.2
Penca. Tronchuda cozinhadas	21.26859771	2.5	2.4	0	0.1	2.4	0.4	0	0.3	0.2	0	0	0	207	58	0.11	1245	71	0.7	26	0.3	2.1
Couve galega cozinhadas	22.94140877	2.9	2.5	0	0.4	2.7	0.4	0	0.3	0.3	0	0.1	0	362	58	0.2	2172	264	0.7	11	0.4	2.1
Brócolos cozinhados	22.46346275	1.3	1	0.2	0.1	2.3	0.7	0.1	0.3	0.1	0	0.1	0	114	18	1.1	687	56	1	12	0.5	2.8
Couve-flor. Couve-bruxelas cozinhada	16.72811056	2.3	1.9	0.1	0.3	1.8	0.2	0	0.1	0.1	0	0	0	5	45	0.11	30	19	0.4	12	0.5	1.6
Grelos. Nabiças. Espinafres cozinhados	16.72811056	1.5	1.3	0.1	0.1	2.3	0.4	0	0.2	0.2	0	0.1	0	161	35	1.1	770	131	0.5	12	0.5	1.9
Feijão verde cozinhado	23.41935478	3.5	2.5	0	1	3	0.3	0	0.2	0.1	0	0.1	0	40	11	0.18	239	41	0.6	17	0.2	1.8
Nabo	13.86043446	2.3	2.2	0	0.1	2.2	0.4	0	0.2	0	0	0	0	3	12	0	20	13	0.2	8	0.1	0.4
Leguminosas cozinhadas: feijão. grão de bico	120.681369	16.7	1	0.6	15.1	5.1	2.1	0.4	1	1	0	0.2	0	4	0	1.1	23	46	2.1	39	1.2	8.4
Ervilha grão. Fava cozinhadas	80.7728767	10.7	1.2	0.4	9.1	5	0.6	0.1	0.2	0.2	0	0.1	0	38	8	0.6	225	56	1.6	38	1	7.9
Salada																						
Alface. Agrião	11.9486504	0.8	0.8	0	0	1.3	0.2	0	0.1	0	0	0	0	115	4	0.6	688	70	1.5	22	0.4	1.8
Cebola	14.8163265	2.4	1.7	0.7	0	1.4	0.2	0	0.1	0.1	0	0	0	0	5	0.15	0	33	0.5	9	0.3	1

Continua na próxima página

Tabela 7.2 Continuação

Item alimentar / Grupo de alimentos	Energia (kcal)	Total HC disponíveis	Mono + dissacarídeos (g)	Oligossacarídeos (g)	Amido (g)	Fibra alimentar (g)	Gordura total (g)	Ácidos gordos monoinsaturados (g)	Ácidos gordos polinsaturados (g)	Ácido linoleico (g)	Ácidos gordos *trans* (g)	Ácidos gordos saturados (g)	Colesterol (mg)	Retinol (Vit.A total) (mg)	Vitamina C (mg)	α-tocoferol (Vit. E) (mg)	Caroteno (mg)	Cálcio (mg)	Ferro (mg)	Magnésio (mg)	Zinco (mg)	Proteína (g)
Cenoura	16.72811056	3.6	3.3	0.1	0.2	3	0	0	0	0	0	0	0	963	2	0.5	5780	45	0.6	6	0.1	0.7
Tomate fresco	19.11784064	3.5	3.5	0	0	1.3	0.3	0.1	0.2	0.2	0	0	0	85	20	1.2	510	11	0.7	11	0.1	0.8
Pimento	30.11059901	3.7	3.5	0.1	0.1	2.8	0.6	0	0.3	0.3	0	0.1	0	383	108	1.4	2300	17	0.9	12	0.3	2.7
Pepino	17.44502958	1.7	1.6	0	0.1	0.7	0.6	0	0.2	0.1	0	0.2	0	6	3	0.07	35	10	0.5	8	0.1	1.4
Fruta																						
Maça. pêra	56.8755759	13.4	13.4	0	0	2.1	0.5	0	0.2	0.1	0	0.1	0	4	7	0.59	26	6	0.2	8	0	0.2
Laranja. Tangerinas	39.90849234	8.7	8.7	0	0	1.7	0.1	0	0.1	0	0	0	0	33	32	0.24	200	30	0.3	9	0.1	0.7
Banana	95.11125718	21.8	19.6	0	2.2	3.1	0.4	0	0.1	0	0	0.1	0	4	10	0.27	21	8	0.4	28	0.2	1.6
Kiwi	53.29098078	10.9	10.9	0	0	1.9	0.5	0.1	0.2	0.1	0	0.1	0	7	72	0.4	42	19	0.4	18	0.2	1.1
Morango	28.67676096	5.3	5.3	0	0	2	0.4	0.1	0.2	0.1	0	0	0	4	47	0.2	26	25	0.8	10	0.1	0.6
Cerejas	60.46017102	13.3	13.3	0	0	1.6	0.7	0.2	0.2	0.2	0	0.2	0	24	6	0.13	141	14	0.4	10	0.1	0.8
Pêssego. Ameixa	37.99670827	8.1	8.1	0	0	2.3	0.3	0.1	0.1	0.1	0	0	0	67	4	0.97	400	8	0.3	8	0.1	0.6
Melão. Melancia	26.7649769	5.7	5.7	0	0	0.9	0.3	0.1	0.1	0.1	0	0.1	0	167	30	0.1	1000	10	0.3	19	0.2	0.6
Diospiro	57.83146794	14.8	14.8	0	0	1.5	0	0	0	0	0	0	0	177	3	0.1	1060	10	0.2	7	0.1	0.6
Figo fresco. Nêsperas. Damascos	69.78011834	16.3	16.3	0	0	2.3	0.5	0.1	0.2	0.2	0	0.1	0	8	1	0.77	50	35	0.6	20	0.1	0.9
Uvas frescas	76.71033557	18.6	18.6	0	0	0.9	0.5	0	0.1	0.1	0	0.1	0	15	1	0.4	60	10	0.3	8	0.1	0.3
Refrigerantes																						
Coca-cola	34.41211315	9	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	1	0	0
Ice-tea	24.37524682	1.5	1.5	0	0	0	0	0	0	0	0	0	0	2	56	0	12	7	0.2	7	0	0.3
Outros refrigerantes. sumos de fruta ou néctares embalados	40.62541136	9.6	9.6	0	0	0.4	0.1	0	0.1	0.1	0	0	0	7	40	0.1	42	6	0.2	4	0.05	0.2
Cerveja e bebidas brancas																						
Cerveja	29.39367998	0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0.3	11	0	0.4
Bebidas brancas: whisky. aguardente. brandy. etc	308.0362073	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Itens alimentares não usados na análise estatística multivariada																						
Frutos conserva: pêssego. ananás	91.28768906	23.2	23.2	0	0	1	0.1	0.1	0	0	0	0	0	2	7	0.06	11	17	0.3	11	0.1	0.2
Frutos secos: amêndoas. avelãs. amendoins. nozes. etc	619.1790637	7.2	4.6	0	2.6	12	56	34.5	14.3	13.9	0	4.7	0	0	1	24	0	266	4	259	3.1	21.6
Azeitonas	172.0605658	0	0	0	0	4	18.5	9.6	2.2	2	0	2.9	0	39	0	2	236	54	1.6	22	0.2	1.4
Chá preto e verde	0.477946016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0.1	0.1
Maionese	657.414745	0	0	0	0	0	71.2	54.6	5.3	4.8	0.006	10.6	122	60	0	10	0	13	0.6	4	0.4	4
Molho de tomate. ketchup	113.2732058	26.9	25.8	0	1.1	1.1	0.3	0.1	0.1	0.1	0	0	0	90	11	1.5	537	13	0.7	19	0.2	1.7
Óleos: girassol. milho. soja	896	0	0	0	0	0	99.5	22.3	62.2	62.1	0.13	11.6	0	0	0	65	0	0	0	0	0	0

Tabela 7.3: Frequência alimentar diária e porção

Frequência alimentar		**Porção**	
Valor atribuído	**Valor da variável (string)**	**Fator atribuído**	**Valor da variável (string)**
0	(*valor omisso*)	0	(*valor omisso*)
1/30	"Nunca ou menos de uma vez por mês"	0.5	"Menor"
2/30	"Uma a três vezes por mês"	1	"Igual"
4/30	"Uma vez por mês"	2	"Maior"
12/30	"Duas a quatro vezes por mês"		
20/30	"Cinco a seis vezes por mês"		
1	"Uma vez por dia"		
2	"Duas a três vezes por dia"		
4	"Quatro a cinco vezes por dia"		
6	"Seis ou mais vezes por dia"		

Tabela 7.4: Descrição das variáveis referentes aos itens alimentares do inquérito e respetivo valor de porção média (g ou ml)

Variável	**Descrição**	**Unidades**	**Porção média padrão (peso edível)**
leiteg	Leite gordo	ml	1 chávena (250 ml)
leitemg	Leite meio-gordo	ml	1 chávena (250 ml)
leitem	Leite magro	ml	1 chávena (250 ml)
iogur	Iogurte	g	1 unidade (125 g)
queij	Queijo (de qualquer tipo, incluindo queijo fresco e requeijão)	g	1 fatia (30 g)
soblact	Sobremesas láteas: pudim, aletria, leite creme	g	1 unidade ou 1 prato de sobremesa (69 g)
gelad	Gelado	g	1 unidade ou 2 bolas (30 g)
ovos	Ovos	g	1 unidade (55 g)
frango	Frango	g	1 porção ou 2 peças (97 g)
perucoe	Perú, coelho	g	1 porção ou 2 peças (100 g)
vacapor	Carnes de vaca, porco, cabrito	g	1 porção (120 g)
figado	Fígado de vaca, porco, frango	g	1 porção (120 g)
lingua	Língua, mão de vaca, tripas, chispe, coração, rim	g	1 porção (100 g)
fiambre	Fiambre, chouriço, salpicão, presunto	g	2 fatias ou 3 rodelas (20 g)
salsich	Salsichas	g	3 médias (75 g)
toucinh	Toucinho, bacon	g	2 fatias (50 g)
peixeg	Peixes gordos: sardinha, cavala, carapau, salmão	g	1 porção (125 g)
peixem	Peixes magros: pescada, faneca, dourada	g	1 porção (125 g)
bacalh	Bacalhau	g	1 porção (125 g)
pconserv	Peixe em conserva: atum, sardinhas	g	1 lata (120 g)
lulapolv	Lulas, polvo	g	1 porção (100 g)
camarao	Camarão, amêijoas, mexilhão	g	1 prato de sobremesa (100 g)
azeite	Azeite	g	1 colher de sopa (10 g)
oleo	Oleos: girassol, milho, soja	g	1 colher de sopa (10 g)
margari	Margarina	g	1 colher de chá (4 g)
manteig	Manteiga	g	1 colher de chá (4 g)
paobran	Pão branco ou tostas	g	1 pão ou 2 tostas (40 g)
paointeg	Pão ou tostas integrais, pão de centeio, pão de mistura	g	1 pão ou 2 tostas (40 g)
broa	Broa, broa de Avintes	g	1 fatia (80 g)
cereais	Flocos de cereais (muesli, corn-flakes, chocapic)	g	1 chávena (40 g)
arroz	Arroz	g	1/2 prato (100 g)
massa	Massas: esparguete, macarrão	g	1/2 prato (100 g)
batfrit	Batatas fritas caseiras	g	1/2 prato (100 g) para caseiras
batfrip	Batatas fritas de pacote	g	1 pacote pequeno (30 g)
batcozi	Batatas cozidas, assadas, estufadas, puré	g	2 batatas médias (160 g)
bolmaria	Bolachas maria, bolachas água e sal, bolachas integrais	g	3 bolachas (21 g)
boloutra	Outras bolachas, biscoitos	g	3 bolachas (21 g)
croipast	Croissants, pastéis, Bolicao, donut, bolos	g	1 unidade ou 1 fatia (80 g)
chocola	Chocolates (tablete ou em pó)	g	3 quadrados para tabletes (18 g) 1 colher de sopa (18 g)
snackcho	Snacks de chocolate (Mars, Twix, Kit-Kat)	g	1 unidade (24 g)
marmcomp	Marmelada, geleia, compota, mel	g	1 colher de sobremesa (15 g)
acucar	Açúcar	g	1 colher de sobremesa ou 1 pacote (9 g)
cbranca	Couve branca, couve lombarda	g	1/2 chávena (75 g)
cpenca	Couve penca, couve tronchuda	g	1/2 chávena (65 g)
cgaleg	Couve galega	g	1/2 chávena (65 g)
brocul	Brocolos	g	1/2 chávena (85 g)
cflor	Couve-flor, couve-bruxelas	g	1/2 chávena (65 g)
grelos	Grelos, nabiças, espinafres	g	1/2 chávena (72 g)
fverde	Feijão verde	g	1/2 chávena (65 g)

Continua na próxima página

Tabela 7.4 Continuação

Variável	Descrição	Unidades	Porção média padrão (peso edível)
alface	Alface, agrião	g	1/2 chávena (15 g)
cebola	Cebola	g	1/2 média (40 g)
cenoura	Cenoura	g	1 média (80 g)
nabo	Nabo	g	1 médio (78 g)
tomate	Tomate fresco	g	1/2 médio (63 g)
piment	Pimento	g	1/2 médio (68 g)
pepino	Pepino	g	1/4 médio (50 g)
legseca	Leguminosas secas: feijão, grão de bico	g	1 chávena (110 g)
ervilha	Ervilhas grão, favas	g	1/2 chávena (70 g)
macaper	Maçã, pera	g	1 média (140 g)
laranja	Laranja, tangerina	g	1 média, 2 médias (153 g)
banana	Banana	g	1 média (100 g)
kiwi	Kiwi	g	1 médio (91 g)
morango	Morangos	g	1 chávena (145 g)
cereja	Cerejas	g	1 chávena (145 g)
pessego	Pêssego, ameixa	g	1 médio, 3 médias (150 g)
melao	Melão, melancia	g	1 fatia média (150 g)
diospir	Diospiro	g	1 médio (200 g)
figond	Figo fresco, nêspera, damasco	g	3 médios (50 g)
uvas	Uvas frescas	g	1 cacho médio (93 g)
fconser	Fruta em conserva (pêssego, ananás)	g	2 metades ou rodelas (63 g)
fseca	Amêndoas, avelãs, amendoins, nozes, pistachios	g	1/2 chávena descascados (35 g)
azeitona	Azeitonas	g	6 unidades (27 g)
vinho	Vinho	ml	1 copo (125 ml)
cerveja	Cerveja	ml	1 garrafa ou 1 lata (330 ml)
bebranca	Bebidas brancas: whisky, aguardente, brandy, vinho do Porto, licores	ml	1 cálice (40 ml)
cola	Coca-Cola, Pepsi, outras colas	ml	1 garrafa ou 1 lata (330 ml)
icetea	Ice-tea	ml	1 garrafa ou 1 lata (330 ml)
refriger	Outros refrigerantes, sumos de fruta, néctares embalados	ml	1 garrafa ou 1 copo (330 ml)
cafe	Café (incluindo pingo, meia de leite e outras bebidas com café)	g	1 chávena de café (50 g)
chapre	Chá preto, chá verde	g	1 chávena (190 g)
croqris	Croquetes, rissóis, bolinhos de bacalhau	g	3 unidades (90 g)
maione	Maionese	ml	1 colher de sobremesa (8 g)
ketchup	Ketchup, molho de tomate	ml	1 colher de sopa (10 g)
pizza	*Pizza*	g	1/2 pizza normal (240 g)
hamburg	Hambúrguer	g	1 médio (100 g)
sopa	Sopa de legumes	ml	1 prato (190 g)

Tabela 7.5: Descrição das variáveis resultantes do agrupamento de itens alimentares

Variáveis dos grupos de alimentos	Variáveis incluídas	Descrição
iogur	iogur	Iogurtes
queij	queij	Queijos
bacalh	bacalh	Bacalhau
azeite	azeite	Azeite
margari	margari	Margarina
manteig	manteig	Manteiga
acucar	acucar	Açúcar
vinho	vinho	Vinho
cafe	cafe	Café
sopa	sopa	Sopa
leite	leiteg + leitemg + leitem	Leite gordo, leite meio-gordo, leite magro
carneb	frango + perucoe	Carne de frango, carne de perú, carne de coelho
carnev	vacapor + figado + lingua	Carne de vaca, carne de porco, cabrito, vísceras
snacks	fiambre + salsich + toucinh + ovos + pizza + hamburg + croqris	Enchidos, salsichas, salgados, pizza, hambúrguer, ovos
peixe	peixeg + peixem	Peixes gordos, peixes magros
moluscos	pconserv + lulapolv + camarao	Lulas, polvo, camarão, amêijoas, mexilhão, pescado em conserva
pao	paobran + paointeg + broa	Pão branco ou tostas, pão integral ou tostas integrais, pão de centeio, pão de mistura, broa, broa de Avintes
flocos	cereais + bolmaria	Flocos de cereais, bolachas maria, bolachas de água e sal e bolachas integrais
acompanhamentos	arroz + massa + batcozi	Arroz, massa, batatas cozidas, assadas, estufadas e puré
batfritas	batfrit + batfrip	Batatas fritas
doces	boloutra + croipast + chocola + snackcho + marmcomp + soblact + gelad	Bolachas doces, croissants ou pastéis, chocolate, marmelada, compotas, sobremesas láteas, gelados
horticolas	cbranca + cpenca + cgaleg + brocul + cflor + grelos + fverde + nabo + legseca + ervilha	Couve branca, couve lombarda, couve penca, couve tronchuda, couve galega, brócolos, couve-flor, couve-bruxelas, grelos, nabiças, espinafres, feijão verde, nabo, leguminosas frescas e secas
salada	alface + cebola + cenoura + tomate + piment + pepino	Alface, agrião, cebola, cenoura, tomate fresco, pepino, pimento
fruta	macaper + laranja + banana + kiwi + morango + cereja + pessego + melao + diospir + figond + uvas	Maçã, pera, laranja, tangerina, banana, kiwi, morangos, cerejas, pêssego, ameixa, melão, melancia, diospiro, figo fresco, nêspera, damasco, uvas frescas, fruta tropical
refrig	cola + icetea + refriger	Coca-Cola, Pepsi, outras colas, outros refrigerantes, sumos de fruta embalados, néctares
bebidasbr	cerveja + bebranca	Cerveja, bebidas brancas como whisky, aguardente, brandy

Tabela 7.6: Descrição das variáveis referentes à composição nutricional dos alimentos

Variável	Descrição INSA
cia_energia_kcal	Calorias kcal. O cálculo da energia utilizou determinados fatores citados em (McCance, 2002).
cia_energia_kj	Calorias kJ. O cálculo da energia utilizou determinados fatores citados em (McCance, 2002).
cia_agua	Água. Corresponde à perda de peso que sofrem os produtos quando aquecidos a temperatura conveniente até peso constante. Inclui a água e uma pequena quantidade de substâncias que se volatilizam nestas condições.
cia_proteina	Proteína. Foi calculada a partir do teor de azoto total, determinado pelo método de Kjeldhal, e multiplicado pelos seguintes fatores (Ferreira et al., 1985): leite e derivados (6.38), trigo e derivados (5.70), restantes alimentos (6.25).
cia_gordura_total	Gordura total
cia_total_hc_disponveis	Total de hidratos de carbono disponíveis. Inclui os monossacarídeos ou açúcares simples (glucose, frutose e galactose), dissacarídeos (sacarose, lactose e maltose), oligossacarídeos (rafinose, estaquiose e verbascose) e polissacarídeos (amido, glicogénio e dextrinas).
cia_total_hc_expresso_monossacar	Total de hidratos de carbono: monossacarídeos expressos
cia_mono_dissacridos	Monossacarídeos, dissacarídeos
cia_acidos_organicos	Ácidos orgânicos
cia_alcool	Álcool
cia_amido	Amido
cia_oligossacaridos	Oligossacarídeos
cia_fibra_alimentar	Fibra alimentar. Inclui polissacáridos não amiláceos solúveis e insolúveis (celulose, pectina e hidrocoloides), lenhina e amido resistente.
cia_acidos_gordos_saturados	Ácidos gordos saturados
cia_acidos_gordos_monoinsaturado	Ácidos gordos monoinsaturados
cia_acidos_gordos_polinsaturados	Ácidos gordos polinsaturados
cia_acidos_gordos_trans	Gorduras *trans*
cia_acido_linoleico	Ácido linoleico
cia_colesterol_mg	Colesterol
cia_retinol_vit_a_total_mg	Total Vitamina A (retinol)
cia_vit_a_total_equivalentes_ret	Vitamina A (equivalentes de actividade de retinol). Corresponde à soma de retinoides e carotenoides com actividade vitamínica, expressos em μg de retinol.
cia_caroteno_mg	Caroteno. O teor de caroteno corresponde ao total dos carotenoides com actividade vitamínica A, expressos em μg de caroteno.
cia_vit_d_microg	Vitamina D. Encontra-se expressa em μg de colecalciferol
cia_atocoferol_mg	Alfa-tocoferol
cia_tiamina_mg	Tiamina
cia_riboflavina_mg	Vitamina B2 (Riboflavina)
cia_equivalentes_de_niacina_mg	Equivalente de niacina
cia_niacina_mg	Niacina
cia_triptofano_60_mg	Triptofano
cia_vit_b6_mg	Vitamina B6. Inclui a soma de piridoxal, piridoxina e piridoxamina e dos seus derivados com fosfato. Encontra-se expressa em mg de piridoxina.
cia_vit_b12_microg	Vitamina B12. Encontra-se expressa em μg de cobalamina.
cia_vit_c_mg	Vitamina C
cia_folatos_microg	Folatos. O valor refere-se a folatos totais. Encontram-se expressos em μg de ácido fólico.
cia_cinza	Cinzas. É o resíduo mineral obtido por incineração, correspondente ao teor de matéria inorgânica presente no alimento.
cia_na_mg	Sódio
cia_k_mg	Potássio
cia_ca_mg	Cálcio
cia_p_mg	Fósforo
cia_mg_mg	Magnésio
cia_fe_mg	Ferro
cia_zn_mg	Zinco

Tabela 7.7: Descrição das variáveis socioeconómicas, sociodemográficas e respeitantes à dieta e ao estilo de vida dos participantes (base de dados original)

	Variável	**Tipo de variável**	**Níveis**	**Descrição**
Demografia				
Idade	`idade`	contínua	-	Idade do participante
Sexo	`sexo`	categórica binária	feminino (`0`), masculino (`1`)	Sexo do participante
Escolaridade				
Escolaridade	`escolaridade`	contínua (n^{o}. inteiro)	-	Anos de escolaridade do participante
Biometria				
Índice de massa corporal	`imc`	contínua	-	Índice de massa corporal do participante (kg/m^2)
Perímetro abdominal	`perimetroabd`	contínua	-	Perímetro abdominal do participante (cm)
Estilo de vida				
Hábitos tabágicos	`fumadorcat`	categórica	não fumador (`0`), fumador (`1`), ex-fumador (`2`)	Hábitos tabágicos do participante
Cigarros por dia	`cigarrosdia`	contínua (n^{o}. inteiro)	-	Cigarros que participante fuma por dia
Anos de tabagismo	`anosfumar`	contínua	-	Anos de tabagismo do participante
Pacotes de tabaco por ano	`pacotesano`	contínua	-	Média de pacotes de tabaco por ano consumidos pelo participante
Horas de actividade física	`horasexercicio`	contínua	-	Horas de exercício realizadas pelo participante (horas/semana)
Consumo energético total	`cia_energia_kcal`	contínua	-	Consumo energético médio do partipante (kcal/dia)
Consumo de álcool	`cia_alcool`	contínua	-	Consumo de álcool do participante (g/dia)
Suplementos	`suplemento`	categórica binária	não (`0`), sim (`1`)	Consumo de suplementos do participante
Desempregado	`desempregado`	categórica binária	não (`0`), sim (`1`)	Estado de desemprego/emprego do participante
Comorbilidades				
Diabetes	`diabetes`	categórica binária	não (`0`), sim (`1`)	Presença de diabetes no participante
Hipertensão	`hipertensao`	categórica binária	não (`0`), sim (`1`)	Presença de hipertensão no participante
Dislipidemia	`dislipidemia`	categórica binária	não (`0`), sim (`1`)	Presença de dislipidemia no participante
Obesidade	`obesidade`	categórica binária	não (`0`), sim (`1`)	Presença de obesidade no participante

Apêndice C - Análises suplementares

Neste anexo são apresentados alguns gráficos, tabelas e a descrição de procedimentos ou análises realizados durante a análise estatística e que complementam alguns dos principais resultados apresentados nos Capítulos 2, 3 e 5.

Com o presente anexo pretende-se, portanto, providenciar uma melhor compreensão e interpretação desses resultados. Este encontra-se dividido nas várias secções e subsecções dos Capítulos ao qual a tabela, gráfico ou procedimento é respeitante, para uma maior facilidade da consulta.

Desenho experimental: Exclusão de participantes

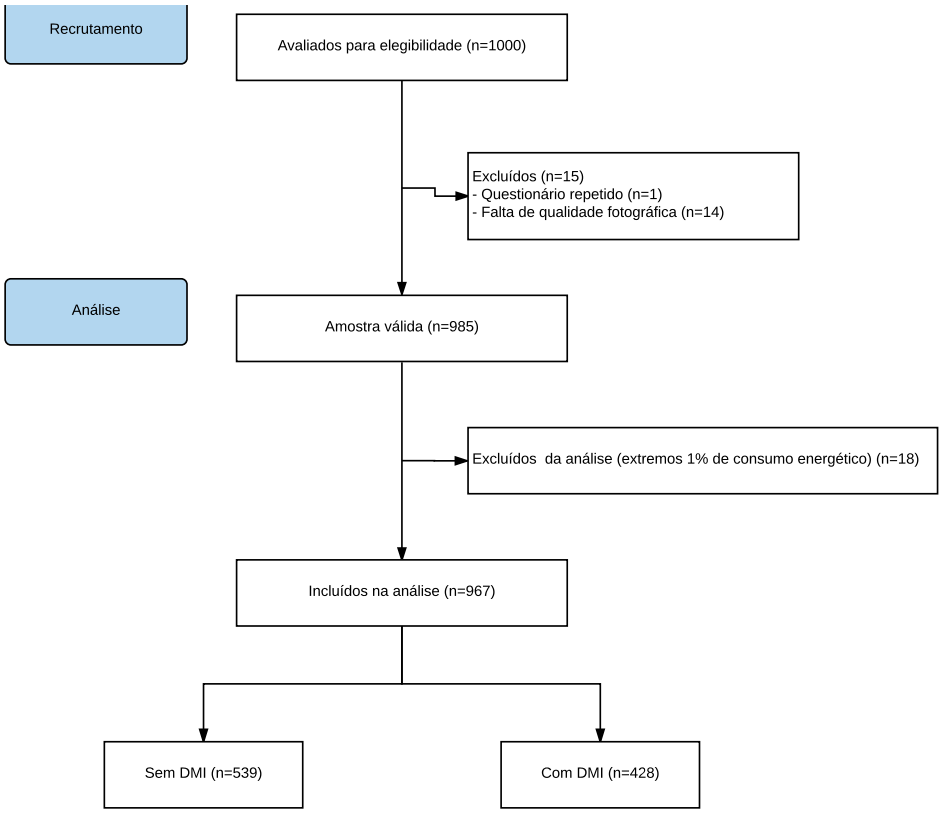
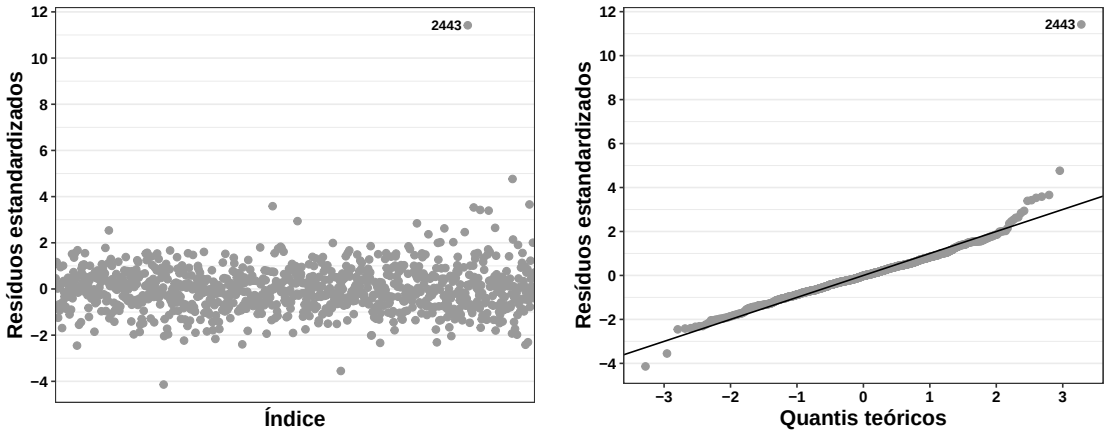


Figura 7.1: Diagrama de fluxo CONSORT dos participantes do estudo selecionados e incluídos na análise

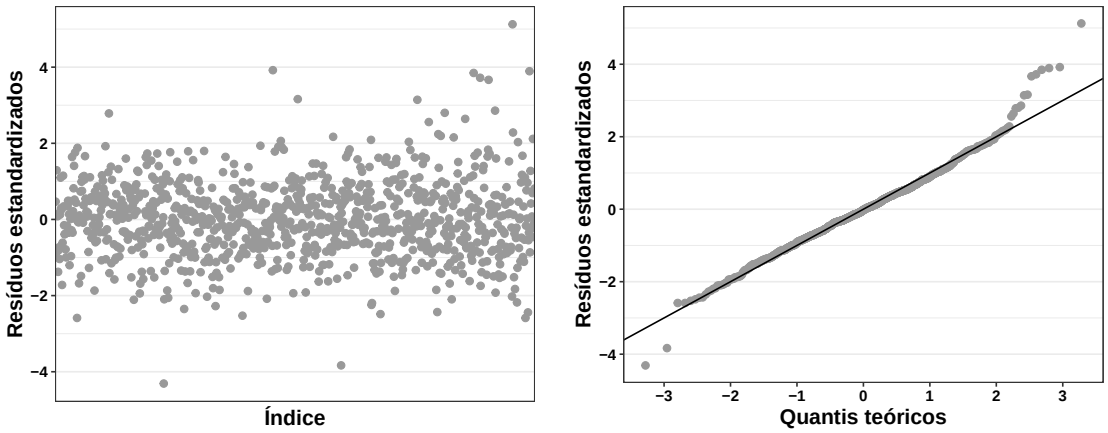
<u>Base de dados:</u> Variáveis referentes às características demográficas dos participantes

- **Imputação de valores**



$E(\texttt{perimetroabd}_i) = 18.751 + 9.127\,\texttt{sexo}_i + 0.169\,\texttt{idade}_i + 2.251\,\texttt{imc}_i$
$R^2 = 0.6406$; AIC= 6800.119; valor-p teste F< 0.001
Valor de alavancagem da obs. 2443 = $0.0150 > \frac{2\times(3+1)}{963}$

Figura 7.2: Gráficos para diagnóstico do modelo linear. À esquerda, gráfico dos resíduos estandardizados por índice do participante. À direita, gráfico quantil-quantil normal (*Q-Q plot*).



$E(\texttt{perimetroabd}_i) = 22.497 + 9.349\,\texttt{sexo}_i + 0.126\,\texttt{idade}_i + 2.218\,\texttt{imc}_i$
$R^2 = 0.6697$; AIC= 6653.462; valor-p teste F< 0.001

Figura 7.3: Gráficos para diagnóstico do modelo linear sem observação de ID 2443. À esquerda, gráfico dos resíduos estandardizados por índice do participante. À direita, gráfico quantil-quantil normal (*Q-Q plot*).

A observação (ID=2443) foi removida pois apresentou ser um *outlier* influente, com valor de alavancagem de 0.0150, superior a $\frac{2\times(3+1)}{963}$, como se pode verificar na figura 7.2 do Apêndice C. De facto, o modelo reajustado sem esse indivíduo apresentou um AIC menor que o do modelo inicial ($\text{AIC}_{inicial} = 6800.119$; $\text{AIC}_{final} = 6653.462$) (ver figura 7.3 do Apêndice C). O modelo reajustado

apresentou também um R^2 de 0.67, o que indica um ajustamento moderadamente bom, e um valor-$p <$ 0.001 no teste F, onde se conclui que existe evidência estatística, aos níveis de significância usuais, de que o modelo reajustado é melhor do que o modelo nulo ($E[\texttt{perimetroabd}] = \beta_0$). Pela análise do gráfico de dispersão dos resíduos do modelo reajustado (gráfico à esquerda, na figura 7.3 do Apêndice C), pode verificar-se que não existe nenhum tipo de padrão que reflita um valor sistematicamente acima (ou abaixo) de zero, o que indica que o valor médio dos resíduos é zero e que esses apresentam variabilidade constante, obedecendo às condições de Gauss-Markov (Gomes, 2011b). A distribuição dos resíduos é aproximadamente normal, à exceção das caudas (gráfico à direita, na figura 7.3 do Apêndice C), o que indica a existência de potenciais *outliers*. No entanto, as observações omissas, de ID 445, 919, 2162 e 3014 não apresentaram valores de idade ou de IMC muito elevados, logo o modelo é aplicável para a respetiva imputação.

Resultados: Identificação de padrões alimentares

- **Análise em componentes principais**

Tabela 7.8: *Loadings* dos grupos alimentares para as componentes principais extraídas sem rotação e com rotação ortogonal *varimax*. Os *loadings* elevados ($| > 0.2|$) encontram-se a negrito, para facilitar a análise da tabela.

	Sem rotação			**Com rotação**		
	CP 1	CP 2	CP 3	CP 1	CP 2	CP 3
Iogurte	-0.0695	**0.4122**	-0.0253	**-0.3225**	**0.2664**	0.0197
Queijo	0.1944	0.1159	0.0297	0.0475	0.1729	0.1414
Bacalhau	**0.2109**	**-0.2130**	-0.1818	**0.3354**	0.0455	-0.0912
Azeite	0.1744	-0.0791	**-0.2547**	**0.2371**	0.1541	-0.1468
Margarina	0.0073	0.0896	0.1314	-0.0886	0.0072	0.1321
Manteiga	0.0928	-0.1150	0.1041	0.1164	-0.0794	0.1131
Açúcar	0.0937	**-0.2199**	0.1992	0.1650	-0.1969	0.1756
Vinho	**0.2420**	**-0.2840**	-0.0910	**0.3827**	-0.0298	-0.0114
Café	0.1807	-0.1433	0.0348	**0.2135**	-0.0206	0.0917
Sopa	0.0561	0.1436	-0.1423	-0.0239	0.1971	-0.0679
Leite	0.0445	0.1673	0.0935	-0.1065	0.0997	0.1321
Carnes brancas	0.1244	-0.0810	0.0503	0.1284	-0.0138	0.0888
Carnes vermelhas	**0.3374**	-0.1632	0.0224	**0.3380**	0.0550	0.1539
Snacks	**0.3328**	0.1033	**0.3176**	0.0800	0.1063	**0.4524**
Peixe (gordo e magro)	0.1674	0.0476	**-0.3535**	0.1705	**0.2855**	**-0.2114**
Moluscos, crustáceos e peixe de conserva	**0.2589**	0.1020	0.0672	0.0922	0.1804	**0.2022**
Pão branco, integral, tostas, broa	0.1894	-0.1745	-0.1149	**0.2779**	0.0306	-0.0376
Cereais, bolachas integrais	0.0410	**0.3040**	0.1366	**-0.2128**	0.1748	0.1921
Arroz, massa, batatas cozidas, assadas	**0.3099**	-0.0805	0.0686	**0.2513**	0.0777	0.1950
Batatas fritas	**0.2632**	-0.0340	**0.3014**	0.1297	-0.0210	**0.3795**
Doces	0.0610	0.0604	**0.3975**	-0.0977	-0.1065	**0.3801**
Hortícolas	**0.2283**	**0.3917**	-0.1142	-0.0814	**0.4524**	0.0854
Salada	**0.2774**	**0.2429**	**-0.2190**	0.0798	**0.4213**	-0.0073
Fruta	**0.2400**	**0.3208**	**-0.2381**	0.0057	**0.4652**	-0.0276
Refrigerantes	0.0809	0.1498	**0.3943**	-0.1441	-0.0311	**0.4034**
Cerveja e bebidas brancas	0.1839	-0.1311	-0.0481	**0.2279**	0.0278	0.0247

• Análise classificatória

Tabela 7.9: Médias de consumo médio diário por cada grupo alimentar e por *cluster* para a solução de 3 *clusters* (métodos *K-means* e *K-medians*)

Alimentos	*K-means*			*K-medians*		
	Cluster 1	***Cluster*** 2	***Cluster*** 3	***Cluster*** 1	***Cluster*** 2	***Cluster*** 3
Iogurte	52.50	29.84	**59.46**	35.10	67.69	**500.00**
Queijo	14.90	**24.81**	16.50	18.71	15.89	2.00
Bacalhau	29.19	**45.97**	32.54	**42.75**	18.00	4.17
Azeite	11.63	**16.88**	13.27	**13.99**	11.81	5.00
Margarina	0.10	0.17	**2.87**	0.48	0.47	0.00
Manteiga	1.64	**1.85**	0.71	**1.71**	1.32	1.07
Açúcar	7.27	**10.06**	9.17	9.39	6.15	2.25
Vinho	63.75	**193.09**	95.03	**138.69**	33.33	0.00
Café	49.99	**79.28**	53.74	**63.34**	48.55	37.50
Sopa	189.25	**219.27**	187.40	196.43	197.40	**228.00**
Leite	237.07	241.53	**301.41**	244.94	249.30	250.00
Carnes brancas	38.72	**49.39**	45.41	47.40	32.99	**52.53**
Carnes vermelhas	27.88	**63.38**	37.13	**47.72**	21.12	11.33
Snacks	19.08	**34.81**	31.04	**28.75**	17.35	8.46
Peixe (gordo e magro)	55.65	**73.80**	48.56	65.29	48.62	**75.00**
Moluscos, crustáceos e peixe de conserva	12.75	**20.12**	17.33	16.40	13.24	0.83
Pão branco, integral, tostas, broa	104.73	**151.42**	97.67	**131.22**	88.36	33.33
Cereais, bolachas integrais	12.39	13.95	15.91	12.89	13.81	**31.20**
Arroz, massa, batatas cozidas, assadas	111.90	**166.11**	123.58	**147.91**	90.43	32.67
Batatas fritas	4.62	**11.50**	7.73	**8.16**	4.35	0.00
Doces	13.39	17.30	**55.11**	20.99	17.47	22.58
Hortícolas	89.48	**135.93**	89.14	110.45	80.07	**953.67**
Salada	81.10	**125.57**	82.76	102.38	74.31	**277.51**
Fruta	363.12	**504.68**	348.45	421.70	351.59	**909.60**
Refrigerantes	18.86	31.78	**60.02**	**30.17**	22.72	5.50
Cerveja e bebidas brancas	5.91	**37.09**	11.11	**19.12**	6.61	5.50

Tabela 7.10: Médias de consumo médio diário por cada grupo alimentar e por *cluster* para a solução de 4 *clusters* (métodos *K-means* e *K-medians*)

Alimentos	*K-means*				*K-medians*			
	Cluster 1	***Cluster 2***	***Cluster 3***	***Cluster 4***	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***
Iogurte	51.40	**96.65**	24.51	47.11	**76.00**	51.30	36.93	32.84
Queijo	24.43	22.46	20.19	14.26	14.15	19.92	20.20	16.38
Bacalhau	30.20	28.18	**52.16**	27.41	16.78	26.65	23.02	**52.97**
Azeite	12.55	**15.54**	**15.88**	11.59	11.25	13.66	11.55	**14.17**
Margarina	**1.10**	0.52	0.34	0.41	0.46	0.55	0.49	0.39
Manteiga	**2.50**	0.84	1.44	1.62	1.13	1.43	1.74	1.88
Açúcar	**11.10**	4.63	**10.83**	7.27	3.45	6.42	**19.87**	9.10
Vinho	87.92	55.02	**227.14**	54.95	25.94	83.56	64.84	**166.78**
Café	62.89	63.11	76.92	47.36	40.30	56.73	**102.26**	55.47
Sopa	196.17	**265.60**	195.00	183.24	186.66	187.96	**323.70**	174.55
Leite	**359.56**	258.50	206.14	242.12	244.24	**267.93**	246.59	225.53
Carnes brancas	**52.31**	39.47	**51.18**	37.02	31.59	46.29	36.04	45.34
Carnes vermelhas	45.03	36.15	**61.48**	26.71	20.45	**52.02**	31.18	34.94
Snacks	**50.62**	27.16	26.60	18.57	16.80	**33.51**	22.28	20.21
Peixe (gordo e magro)	53.57	**76.91**	**69.17**	52.31	43.85	**68.91**	62.10	56.56
Moluscos, crustáceos e peixe de conserva	21.31	17.93	17.64	12.47	12.02	**18.85**	13.23	13.72
Pão branco, integral, tostas, broa	113.78	110.78	**149.36**	101.87	81.78	121.49	**130.49**	122.71
Cereais, bolachas integrais	**25.70**	18.23	9.84	11.51	13.62	**16.35**	12.84	10.01
Arroz, massa, batatas cozidas, assadas	**174.34**	123.46	153.90	107.46	76.96	**151.58**	132.47	125.82
Batatas fritas	**19.16**	5.63	7.40	4.52	4.23	**9.54**	5.98	5.49
Doces	**51.48**	14.66	16.30	16.65	16.96	19.75	**26.34**	19.24
Hortícolas	118.44	**240.66**	93.90	72.54	67.86	**141.63**	87.39	80.71
Salada	95.84	**181.40**	94.65	72.79	64.71	**123.01**	82.31	78.64
Fruta	400.84	**656.52**	388.38	347.44	327.18	**500.10**	349.72	341.60
Refrigerantes	**108.45**	27.73	17.92	17.13	21.29	**36.12**	24.14	22.63
Cerveja e bebidas brancas	10.77	9.12	**39.26**	5.11	3.10	**22.14**	6.47	15.03

- **Comparação dos padrões obtidos com os dois métodos**

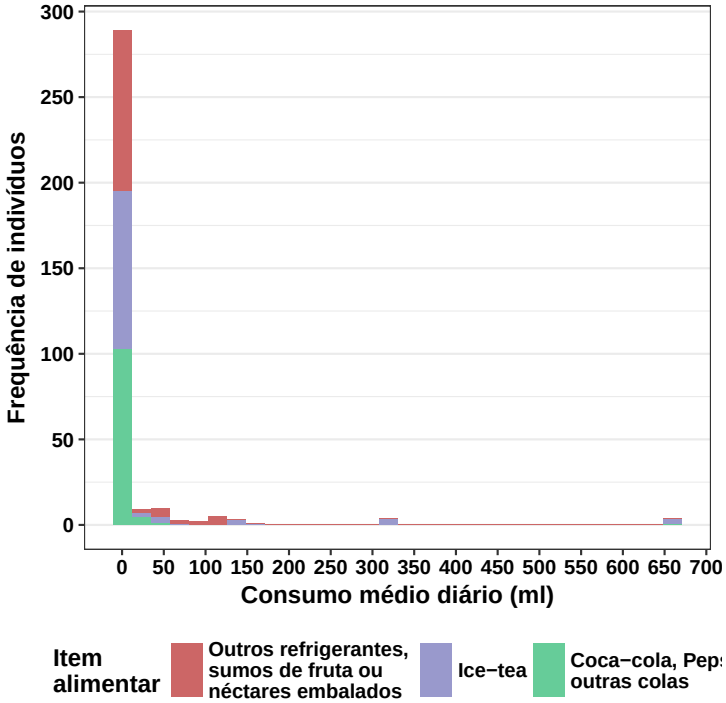


Figura 7.4: Histograma do consumo médio diário dos itens alimentares que compõem a variável `refriger` para os indivíduos do *Cluster* 3

• **Características dos indivíduos por padrão alimentar**

Tabela 7.11: Estatísticas descritivas dos *scores* para cada componente principal (CP): mediana, amplitude interquartil (AIQ), mínimo (mín), máximo (máx) e valores do 1º e 3º tercil

	Mediana	**AIQ**	**Mín**	**Máx**	**1º Tercil**	**2º Tercil**
CP 1	-0.091	1.217	-4.005	3.75	-0.472	0.315
CP 2	-0.18	1.109	-2.206	5.915	-0.505	0.185
CP 3	-0.191	1.096	-2.47	8.812	-0.49	0.174

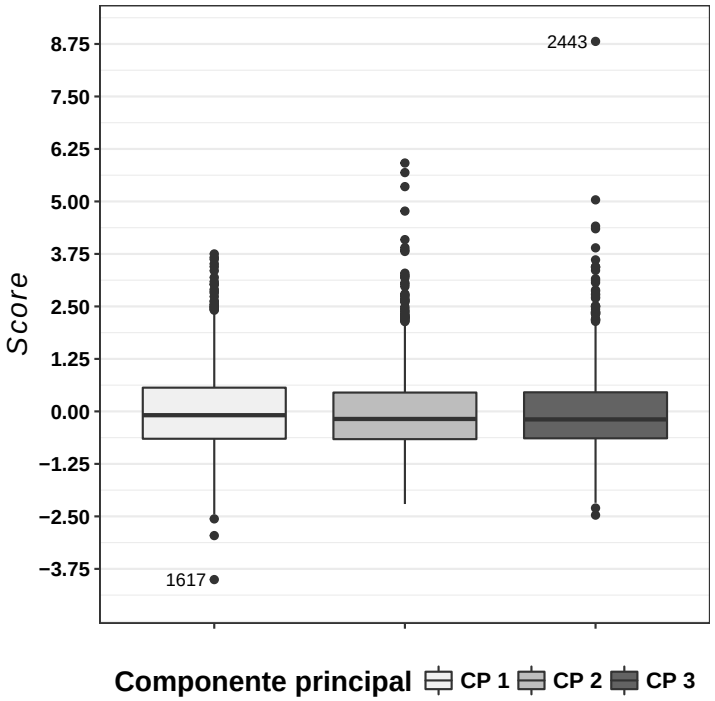
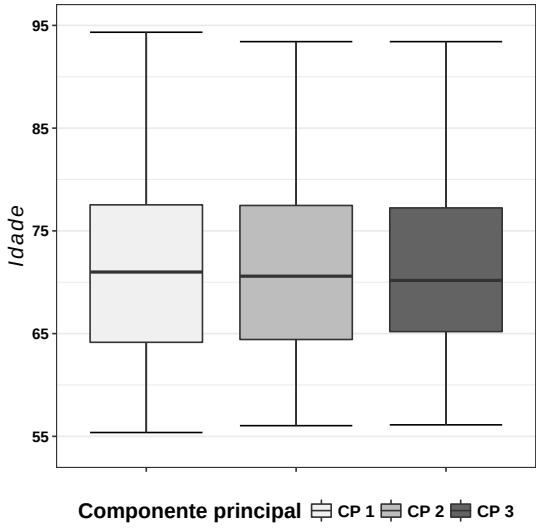
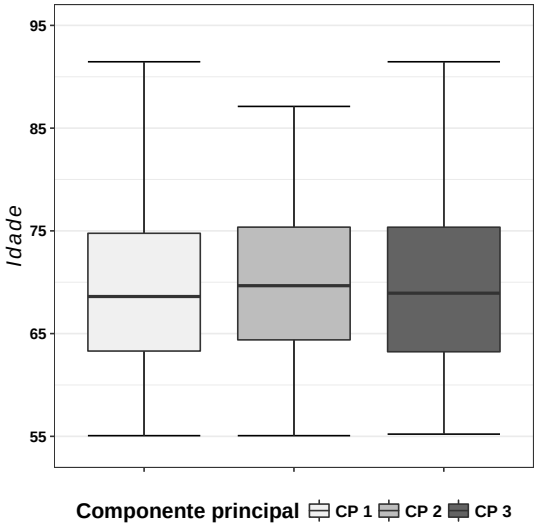


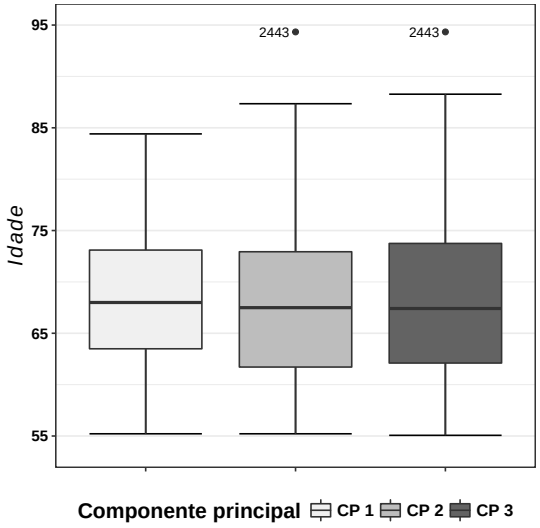
Figura 7.5: Boxplot dos *scores* para cada componente principal (CP)



(a) Boxplot da distribuição da idade por componente principal (CP): Tercil 1



(b) Boxplot da distribuição da idade por componente principal (CP): Tercil 2



(c) Boxplot da distribuição da idade por componente principal (CP): Tercil 3

Figura 7.6: Gráficos boxplot da distribuição da idade por tercil de cada componente principal (CP)

Estandardização de variáveis pela idade

O primeiro passo na estandardização por idade consistiu na decisão de qual o tercil de cada componente principal a utilizar como referência, onde neste projeto foi escolhido o primeiro tercil (colunas a negrito na tabela 7.12 abaixo apresentada). As proporções de cada tercil de referência dentro de cada componente principal k ($k = 1, 2, 3$) foram definidas por r_{ik}, com i ($i = 1, 2, 3$) representando o índice de cada classe etária, por ordem crescente de grupos de idade.

Tabela 7.12: Distribuição dos participantes por grupo etário e por tercil de cada componente principal

	Componente principal 1			**Componente principal 2**			**Componente principal 3**		
Faixa etária	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
"55 - 64 anos"	**80 (0.25)**	90 (0.28)	91 (0.28)	**74 (0.23)**	76 (0.24)	111 (0.34)	**57 (0.18)**	92 (0.29)	112 (0.35)
"65 - 74 anos"	**116 (0.36)**	143 (0.44)	157 (0.49)	**129 (0.40)**	143 (0.44)	144 (0.45)	**150 (0.46)**	132 (0.41)	134 (0.42)
"$\geq$ 75 anos"	**127 (0.39)**	89 (0.28)	74 (0.23)	**120 (0.37)**	103 (0.32)	67 (0.21)	**116 (0.36)**	98 (0.30)	76 (0.24)
valor-*p* [a]	<0.001***			<0.001***			<0.001***		

[a] Valores-p de testes de qui-quadrado de homogeneidade entre os vários tercis de cada componente principal.
*** Significância estatística ao nível α=0.001

O valor estandardizado para cada tercil j ($j = 1, 2, 3$) de cada componente principal k nas variáveis categóricas foi assim obtido através do cálculo $\sum_{i=1}^{3} p_{ijk}\, r_{ik}$, onde p_{ijk} é a proporção de indivíduos no nível de interesse da variável categórica em estudo, para a classe etária i, tercil j e componente principal k. Já nas variáveis contínuas, foram primeiramente determinadas as médias da variável em questão ($\bar{x}_{ijk}$) dentro de cada faixa etária, tercil e componente principal, e efetuado o cálculo $\sum_{i=1}^{3} \bar{x}_{ijk}\, r_{ik}$.

Na tabela 7.13 são apresentados os cálculos efetuados na estandardização das variáveis categóricas e contínuas para melhor compreensão do processo.

Tabela 7.13: Cálculos efetuados na estandardização das variáveis categóricas e contínuas.

	Variável categórica			**Variável contínua**		
	CP 1	CP 2	CP 3	CP 1	CP 2	CP 3
Tercil 1	$\sum_{i=1}^{3} p_{i11}\, r_{i1}$	$\sum_{i=1}^{3} p_{i12}\, r_{i2}$	$\sum_{i=1}^{3} p_{i13}\, r_{i3}$	$\sum_{i=1}^{3} \bar{x}_{i11}\, r_{i1}$	$\sum_{i=1}^{3} \bar{x}_{i12}\, r_{i2}$	$\sum_{i=1}^{3} \bar{x}_{i13}\, r_{i3}$
Tercil 2	$\sum_{i=1}^{3} p_{i21}\, r_{i1}$	$\sum_{i=1}^{3} p_{i22}\, r_{i2}$	$\sum_{i=1}^{3} p_{i23}\, r_{i3}$	$\sum_{i=1}^{3} \bar{x}_{i21}\, r_{i1}$	$\sum_{i=1}^{3} \bar{x}_{i22}\, r_{i2}$	$\sum_{i=1}^{3} \bar{x}_{i23}\, r_{i3}$
Tercil 3	$\sum_{i=1}^{3} p_{i31}\, r_{i1}$	$\sum_{i=1}^{3} p_{i32}\, r_{i2}$	$\sum_{i=1}^{3} p_{i33}\, r_{i3}$	$\sum_{i=1}^{3} \bar{x}_{i31}\, r_{i1}$	$\sum_{i=1}^{3} \bar{x}_{i32}\, r_{i2}$	$\sum_{i=1}^{3} \bar{x}_{i33}\, r_{i3}$

Resultados: Regressão logística binária

- **Seleção de covariáveis**

Tabela 7.14: Processo de seleção de variáveis do modelo por eliminação *backward*

Passo 1		Passo 2		Passo 3	
`sexo`		`diabetes`		`imc`	
valor-p RV	$\Delta\beta_{sexo}$	valor-p RV	$\Delta\beta_{diabetes}$	valor-p RV	$\Delta\beta_{imc}$
0.562	`pc1`:3.613	0.570	`pc1`:1.833	0.544	`pc1`:2.228
	`pc2`:0.310		`pc2`:4.388		`pc2`:3.066
	`pc3`:43.822		`pc3`:6.976		`pc4`:23.484
		Novo modelo sem `diabetes`			
`esc_cat`				`sexo`	
valor-p RV	$\Delta\beta_{esc\_cat}$			valor-p RV	$\Delta\beta_{sexo}$
0.553	`pc1`:1.542			0.480	`pc1`:2.746
	`pc2`:0.913				`pc2`:4.351
	`pc3`:8.334				`pc3`:44.581
Novo modelo sem `esc_cat`					
				`idade`	
				valor-p RV	$\Delta\beta_{idade}$
				0.437	`pc1`:1.618
					`pc2`:5.197
					`pc3`:82.601
				`cia_energia_kcal`	
				valor-p RV	$\Delta\beta_{cia\_energia\_kcal}$
				0.428	`pc1`:23.270
					`pc2`:31.392
					`pc3`:481.184
				`dislipidemia`	
				valor-p RV	$\Delta\beta_{dislipidemia}$
				0.351	`pc1`:2.478
					`pc2`:3.186
					`pc3`:24.633
				`exercicio`	
				valor-p RV	$\Delta\beta_{exercicio}$
				0.288	`pc1`:2.248
					`pc2`:10.134
					`pc3`:12.726
				Mais nenhuma variável com valor-p>0.20	

Tabela 7.15: Valores de AIC e testes de não linearidade para modelo com diferentes formas das variáveis de confundimento

Transformação	`imc` (valor-p=0.544)		`idade` (valor-p=0.470)		`cia_energia_kcal` (valor-p=0.433)		`perimetroabd` (valor-p=0.237)		`alcool` (valor-p=0.161)		`pacotesano` (valor-p=0.068)	
	AIC	**valor-p RV**	**AIC**	**valor-p RV**	**AIC**	**valor-p RV**	**AIC**	**valor-p RV**	**AIC**	**valor-p RV**	**AIC**	**valor-p RV**
Linear	1332.956	-	1330.500	-	1327.384	-	1324.423	-	1325.511	-	1326.522	-
Restricted cubic splines (Nº. nós)												
3	1334.585	0.543	1332.447	0.818	1327.727	0.198	1324.299	0.145	1323.239	0.039**	1328.461	0.805
4	1334.654	0.316	1332.580	0.383	1329.118	0.322	1325.591	0.243	-	-	1326.107	0.110
5	1336.263	0.441	1332.956	0.315	1330.500	0.410	1327.384	0.386	1324.423	0.079*	1323.239	0.026**
6	1331.622	0.053*	1334.347	0.386	1331.916	0.483	1328.255	0.384	1324.523	0.072*	1326.052	0.076*
7	1333.945	0.109	1336.359	0.529	1328.203	0.102	1327.851	0.255	-	-	-	-
8	-	-	1337.073	0.490	1330.969	0.209	1329.532	0.331	-	-	-	-
9	-	-	1338.381	0.526	1331.039	0.170	1331.575	0.445	-	-	-	-
Polinomial												
2º grau	1334.931	0.874	1332.282	0.641	1328.989	0.530	1324.819	0.205	1324.617	0.089*	1328.412	0.740
3ºgrau	1333.312	0.162	1333.293	0.547	1327.952	0.180	1324.985	0.179	1326.596	0.233	1329.921	0.740

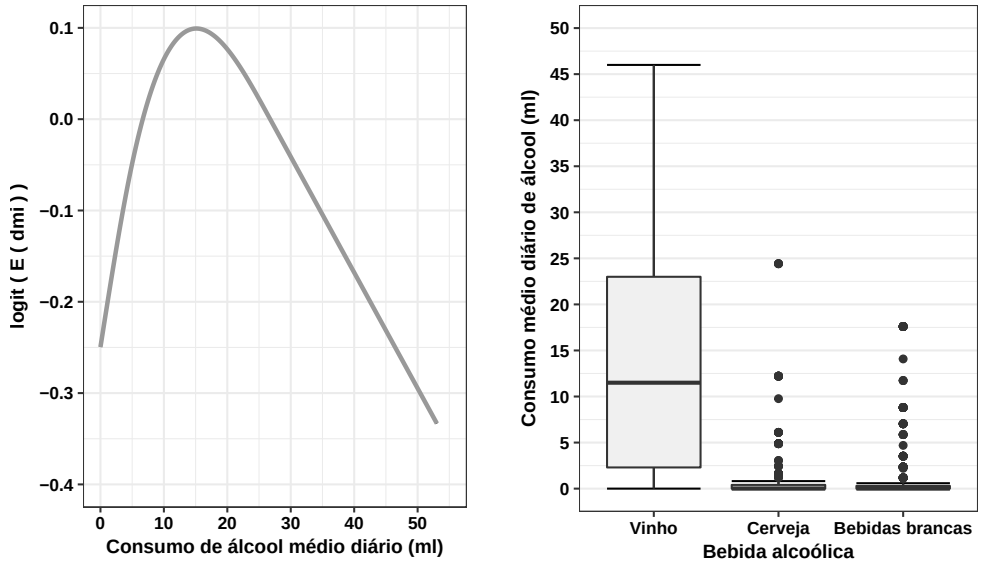
Significância estatística aos níveis α=0.1 (*) e α=0.05 (**)

Tabela 7.16: Valores-p dos testes da razão de verosimilhanças (RV) entre o *Modelo II* (sem interações) e o *Modelo II* com a interação, e valores de AIC e graus de liberdade (g.l.) para cada modelo

Interação	**g.l.**	**AIC**	**Valor-p RV**	Interação (continuação)	**g.l.**	**AIC**	**Valor-p RV**
Sem interações	-	1323.2	-	idade:hipertensao	1	1323.9	0.2547
pc1:pc2	1	1324.7	0.4566	cia_energia_kcal:exercicio	1	1323.9	0.2477
pc1:pc3	1	1325.1	0.6889	cia_energia_kcal:sexo	1	1325.2	0.7872
pc1:idade	1	1323.8	0.2227	cia_energia_kcal:perimetroabd	1	1322.2	0.0832*
pc1:cia_energia_kcal	1	1323.8	0.2301	cia_energia_kcal:imc	1	1325.2	0.8404
pc1:exercicio	1	1325.2	0.8670	cia_energia_kcal:dislipidemia	1	1323.1	0.1417
pc1:sexo	1	1325	0.6089	cia_energia_kcal:$S_{(A,3)}$	2	1323.1	0.1261
pc1:perimetroabd	1	1325.2	0.9313	cia_energia_kcal:$S_{(P,5)}$	4	1325	0.1816
pc1:imc	1	1323.3	0.1610	cia_energia_kcal:jafumador	1	1325.2	0.8748
pc1:dislipidemia	1	1323.4	0.1746	cia_energia_kcal:hipertensao	1	1322.6	0.1056
pc1:$S_{(A,3)}$	2	1327	0.8932	exercicio:sexo	1	1325.2	0.9595
pc1:$S_{(P,5)}$	4	1330.1	0.8877	exercicio:perimetroabd	1	1325.2	0.9303
pc1:jafumador	1	1324.5	0.4009	exercicio:imc	1	1323	0.1377
pc1:hipertensao	1	1324.5	0.3798	exercicio:dislipidemia	1	1325.2	0.8399
pc2:pc3	1	1325	0.6631	exercicio:$S_{(A,3)}$	2	1324.2	0.2140
pc2:idade	1	1324.9	0.5546	exercicio:$S_{(P,5)}$	4	1326.4	0.3040
pc2:cia_energia_kcal	1	1325	0.6378	exercicio:jafumador	1	1324.4	0.3553
pc2:exercicio	1	1325.1	0.7153	exercicio:hipertensao	1	1324	0.2747
pc2:sexo	1	1325.1	0.7349	sexo:perimetroabd	1	1321	0.0390**
pc2:perimetroabd	1	1324.8	0.4943	sexo:imc	1	1324.1	0.2846
pc2:imc	1	1325.1	0.7165	sexo:dislipidemia	1	1325.2	0.9966
pc2:dislipidemia	1	1325.1	0.6997	sexo:$S_{(A,3)}$	2	1327	0.9034
pc2:$S_{(A,3)}$	2	1321.7	0.0619*	sexo:$S_{(P,5)}$	4	1323.3	0.0953*
pc2:$S_{(P,5)}$	4	1320.4	0.0281**	sexo:jafumador	1	1324.1	0.2781
pc2:jafumador	1	1324	0.2699	sexo:hipertensao	1	1322.4	0.0928*
pc2:hipertensao	1	1325	0.6454	perimetroabd:imc	1	1324	0.2692
pc3:idade	1	1324.9	0.5785	perimetroabd:dislipidemia	1	1323.9	0.2417
pc3:cia_energia_kcal	1	1325	0.6332	perimetroabd:$S_{(A,3)}$	2	1326	0.5339
pc3:exercicio	1	1324.6	0.4143	perimetroabd:rcs(pacotesano,5)	4	1328.4	0.5807
pc3:sexo	1	1325.1	0.7239	perimetroabd:jafumador	1	1325.2	0.8332
pc3:perimetroabd	1	1322	0.0706*	perimetroabd:hipertensao	1	1325.2	0.8974
pc3:imc	1	1323.6	0.2035	imc:dislipidemia	1	1325.2	0.7986
pc3:dislipidemia	1	1325.2	0.9191	imc:$S_{(A,3)}$	2	1325.3	0.3809
pc3:$S_{(A,3)}$	2	1324.8	0.2897	imc:$S_{(P,5)}$	4	1327.8	0.4807
pc3:$S_{(P,5)}$	4	1325.3	0.2078	imc:jafumador	1	1325	0.6006
pc3:jafumador	1	1324.8	0.4858	imc:hipertensao	1	1324.6	0.4087
pc3:hipertensao	1	1322.1	0.0755*	dislipidemia:$S_{(A,3)}$	2	1326.5	0.6964
idade:cia_energia_kcal	1	1325.2	0.8505	dislipidemia:$S_{(P,5)}$	4	1329.5	0.7827
idade:exercicio	1	1320.1	0.0238**	dislipidemia:jafumador	1	1325.2	0.7768
idade:sexo	1	1325	0.6328	dislipidemia:hipertensao	1	1322.5	0.0956*
idade:perimetroabd	1	1325.2	0.9086	$S_{(A,3)}$:$S_{(P,5)}$	8	1333.3	0.6599
idade:imc	1	1324.7	0.4707	$S_{(A,3)}$:jafumador	2	1323.6	0.1653
idade:dislipidemia	1	1324.7	0.4717	$S_{(A,3)}$:hipertensao	2	1325.5	0.4247
idade:$S_{(A,3)}$	2	1326.4	0.6699	$S_{(P,5)}$:hipertensao	4	1328.3	0.5682
idade:$S_{(P,5)}$	4	1329.7	0.8178	jafumador:hipertensao	1	1325	0.6004
idade:jafumador	1	1324.8	0.4858				

Significância estatística ao níveis $\alpha = 0.1$ (*) e $\alpha = 0.05$ (**)

$S_{(A,3)}$ e $S_{(P,5)}$ correspondem às variáveis alcool e pacotesano e componentes RCS associadas de 3 e 5 nós, respetivamente



(a) Gráfico dos valores *logit* estimados para a variável `alcool` e componentes RCS associadas

(b) Boxplot da distribuição do consumo de álcool médio diário segundo diferentes bebidas alcoólicas

Figura 7.7: Gráficos referentes ao consumo alcoólico e seu impacto no risco de DMI (sem observações de ID 2443, 1268, 967 e 1365)

Apêndice D - Exemplo de código

Principais funções e bibliotecas usadas

Tabela 7.17: Principais funções e bibliotecas usadas

Análise estatística	Biblioteca	Funções	Referência
Análise em componentes principais		`[MV]pca`	(StataCorp, 2013b)
Análise classificatória		`[MV]cluster`	(StataCorp, 2013b)
Modelo linear	`stats`	`lm()`	(R Core Team and contributors worldwide, 2017)
Modelos de regressão logística binária com *restricted cubic splines* e/ou polinómios	`rms`	`lrm()`	(Harrell Jr, 2017)
Multicolinearidade	`car`	`vif()`	(Fox and Weisberg, 2016)
Análise de resíduos		`outlierTest()`	
		`leveragePlots()`	
		`influencePlot()`	
Construção de gráficos	`ggplot2`	`ggplot()`	(Wickham and Chang, 2016)
	`gridExtra`	`grid.arrange()`	(Auguie and Antonov, 2016)

Exemplos de código

Código Stata - Análise em componentes principais

```
*******************************************************************************
*Dados
*******************************************************************************
use "C:\Users\AsusPC\Desktop\AIBILI\Database INSA\
20160106_IA_Database_gruposalimentos_exclusao.dta",clear

*******************************************************************************
*Análise em componentes principais
*->Lembrar que vamos usar matriz de correlações -> dados estandardizados
*******************************************************************************

* Alterar diretório
local path "C:\Users\AsusPC\Desktop\AIBILI\Analise exploratoria\Caracterizacao Alimentos\PCA"
cd "`path'"
```

```
* Análise em componentes principais
pca iogur queij bacalh azeite margari manteig acucar vinho cafe sopa leite carneb
carnev snacks peixe moluscos pao flocos acompanhamentos batfritas doces horticolas salada fruta
refrig bebidasbr

* Scree-plot
screeplot, yline(1)
graph export "Screeplotalimentos_sem_aj.png", replace * exportar gráfico

* Rotação ortogonal varimax das 3 primeiras CP's
rotate, comp(3)

* Exportar resultados para txt
translate @Results PCA_grupos_rotacao.txt

* Criar variáveis dos scores para as componentes principais extraídas
predict pc1 pc2 pc3

* Guardar base de dados
save "C:\Users\AsusPC\Desktop\AIBILI\Database INSA\20160106_IA_Database_pca_scores.dta",replace

clear all
```

Código R - Regressão logística binária (processo de modelação)

```
library(gridExtra)
library(car)
library(ggplot2)
library(readstata13)
library(rms)
library(stats)

#############################################################################
#Funções de apoio
#############################################################################

## Função para obter gráfico lowess
logitloess <- function(x, y, s) {
logit <- function(pr) {
log(pr/(1-pr))
}
if (missing(s)) {
locspan <- 0.7
} else {
locspan <- s
}
loessfit <- predict(loess(y~x,span=locspan))
pi <- pmax(pmin(loessfit,0.9999),0.0001)
logitfitted <- logit(pi)
plot(x, logitfitted, ylab="logit")
}

#############################################################################
##Dados
#############################################################################
```

```
data<-read.table('F:/Tese Final/AIBILI/Do-files/pca_data.txt',h=T,sep='')
rownames(data)<-data$id
names(data)<- tolower(names(data))

##Fatorização e renomeação das variáveis dmi, sexo,jafumador,diabetes, suplemento, hipertensao,
dislipidemia, exercicio, escolaridade

##Criar variável alcool

##Excluir variáveis

## Formato dos dados para criar modelos com pacote rms d<-data
dd<-datadist (d); options (datadist ="dd")
dmi_num<- as.numeric(data$dmi)-1 #transformar variável dmi em numérica para construção de
gráficos

#############################################################################
#1) GVIF - Estudo da multicolinearidade
#############################################################################

table<- NULL
fit1 = glm(dmi ~ pc1+pc2+pc3+ idade+ imc +cia_energia_kcal + sexo+perimetroabd +exercicio+
diabetes+ hipertensao+ dislipidemia +alcool+ jafumador+pacotesano+esc_cat, family =
binomial(logit), data = data)
vifs<- vif(fit1)
num<- seq(1:17)
table<- cbind(num, vif(fit1))
round(table,3)
#Nenhuma variável com VIF>5

#############################################################################
#2) Análise univariada
#############################################################################

## a) Categóricas -> sexo, exercicio, jafumador, diabetes,hipertensao, dislipidemia, esc_cat
final<- NULL
##sexo
(tabsexo<- table(data$sexo, data$dmi,dnn=c('Sexo','DMRI')))
summary(glm(dmi~sexo,binomial,data))
final<- rbind(final, summary(glm(dmi~sexo,binomial,data))$coef)

## b) Tabela univariada

## c) Teste da razão de verosimilhanças
nomes=names(data) %in% c('dmi',"id","jafumador","pacotesano")
nomes2<-names(data[,!nomes])
table2=NULL
for (nome in nomes2){
table2<- rbind(table2, round(anova(glm(dmi~1,binomial,data),glm(dmi~eval(parse(text = nome)),
binomial,data),test='Chisq')[2,5],3))
}
table2<- cbind(nomes2,table2)
table2=as.matrix(table2)
colnames(table2)<-c('Variable','p-val')
table2<-table2[order(as.numeric(table2[,2])) , ]
round(anova(glm(dmi~1,binomial,data),glm(dmi~jafumador+pacotesano,
binomial,data),test='Chisq')[2,5],3)
```

```
##############################################################################
#3) Construção de modelo múltiplo
##############################################################################

##Modelo completo
form=dmi∼rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,5)+rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5) +
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat
mod1=glm(form,binomial,data)
round(summary(mod1)$coef,4)

##Modelo completo com função lrm do pacote rms
f<-lrm( dmi∼rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,5)+ rcs(imc,5)+rcs(idade,5) +rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat, data =d, x=TRUE, y=TRUE)

########## a) Estudar linearidade das variáveis de interesse
RV=NULL
RV<-rbind(RV,anova(update(mod1,∼.  -rcs(pc1,5)), mod1,test="Chisq")[2,5] )
RV<-rbind(RV,anova(update(mod1,∼.  -rcs(pc2,5)), mod1,test="Chisq")[2,5] )
RV<-rbind(RV,anova(update(mod1,∼.  -rcs(pc3,5)), mod1,test="Chisq")[2,5] )
rownames(RV)<- c("pc1", "pc2","pc3")
RV<-RV[order(as.numeric(RV[,1])) , ]
RV<- cbind(seq(1:3), RV)

## Análise do gráfico lowess da variável pc3
logitloess(data$pc3, dmi_num)
plot(data$pc3)
ind<- which(data$pc3==max(data$pc3))
data[which(data$pc3==max(data$pc3)),] #id=1268
text(ind,data$pc3[ind],labels =data$id[ind], pos=4)

## Modelos com variável pc3 na forma restricted cubic splines-> 2 a 9 nós _aic<- 0
b_aic<-0
c_aic<-0
d_aic<-0
e_aic<-0
f_aic<-0
g_aic<-0
h_aic<-0
tab<-NULL
a_aic<- lr(dmi∼rcs(pc1,5)+rcs(pc2,5)+pc3+rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5) +sexo+
rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat)
b_aic<- lr(dmi∼rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,3)+rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat)
c_aic<-lr(dmi∼rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,4)+rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat)
d_aic<- lr(dmi∼rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,5)+ rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat)
e_aic<- lr(dmi∼rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,6)+ rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat)
f_aic<- lr(dmi∼rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,7)+rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+ dislipidemia+diabetes+rcs(alcool,5)+
```

```
rcs(pacotesano,5)+jafumador+esc_cat)
g_aic<- lr(dmi~rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,8)+ rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat)
h_aic<- lr(dmi~rcs(pc1,5)+rcs(pc2,5)+rcs(pc3,9)+ rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat)
knots<- c(0,3,4,5,6,7,8,9)
AIC<-c(AIC(a_aic),AIC(b_aic),AIC(c_aic),AIC(d_aic),AIC(e_aic),AIC(f_aic),AIC(g_aic),AIC(h_aic))
(tab<-cbind(knots,AIC))

##Polinómios 2 e 3 graus
pol2<- lrm(dmi~rcs(pc1,5)+rcs(pc2,5)+poly(pc3,2)+ rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat, data =d, x=TRUE, y=TRUE)
pol3<- lrm(dmi~rcs(pc1,5)+rcs(pc2,5)+poly(pc3,3)+ rcs(imc,5)+rcs(idade,5)+rcs(cia_energia_kcal,5)+
sexo+rcs(perimetroabd,5)+exercicio+hipertensao+dislipidemia+diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat, data =d, x=TRUE, y=TRUE)

##Testar linearidade da variável contínua - Teste RV modelo linear vs.  modelos não lineares
lrtest (a_aic,b_aic) # -> linear
lrtest (a_aic,c_aic)
lrtest (a_aic,d_aic)
lrtest (a_aic,e_aic)
lrtest (a_aic,f_aic)
lrtest (a_aic,g_aic)
lrtest (a_aic,h_aic)
lrtest (a_aic,pol2)
lrtest (a_aic,pol3)

##Selecionar a variável de interesse com menor significância no novo modelo reajustado para
testar linearidade
RV=NULL
RV<-rbind(RV,anova(update(mod2,~.  -rcs(pc1,5)), mod2,test="Chisq")[2,5] )
RV<-rbind(RV,anova(update(mod2,~.  -rcs(pc2,5)), mod2,test="Chisq")[2,5] )
rownames(RV)<- c("pc1", "pc2")
RV<-RV[order(as.numeric(RV[,1])) , ]
RV<- cbind(seq(1:2), RV)

## Repetição procedimento para as restantes variáveis de interesse

## Modelo obtido
form=dmi~pc1+pc2+pc3+ rcs(imc,5)+rcs(idade,5) +rcs(cia_energia_kcal,5) + sexo+
rcs(perimetroabd,5) +exercicio+hipertensao+ dislipidemia +diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat
mod4=glm(form,binomial,data)
round(summary(mod4)$coef,4)
f4<-lrm( dmi~pc1+pc2+pc3+ rcs(imc,5)+rcs(idade,5) +rcs(cia_energia_kcal,5) +
sexo+rcs(perimetroabd,5) +exercicio+hipertensao+ dislipidemia +diabetes+rcs(alcool,5)+
rcs(pacotesano,5)+jafumador+esc_cat , data =d, x=TRUE, y=TRUE)

########## b) Seleção variáveis de confundimento - critério de alteração dos β vs α ≥ 0.2

## Teste RV - modelo corrente vs modelo sem variável

#1) Tirar esc_cat
m1<- update(mod4,~.  -esc_cat) #
```

```
## Teste RV - modelo corrente vs modelo sem variável

#2) tirar diabetes
m2<- update(m1,∼.  -diabetes)

## Teste RV - modelo corrente vs modelo sem variável

#3) Tirar imc
m3<- update(m2,∼.  -rcs(imc,5))

## Teste RV - modelo corrente vs modelo sem variável

#Não tirar mais nenhuma variável

########## c) Testar linearidade das variáveis de confundimento

###############################################################################
#4) Avaliação e diagnóstico do modelo ajustado (Modelo I)
###############################################################################
```