Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Seleção de variáveis na presença de valores omissos: uma aplicação na modelação do Índice de Massa Corporal nos imigrantes africanos e brasileiros residentes em Lisboa e Setúbal

Beatriz Goulão

Trabalho de Projeto

Mestrado em Bioestatística

2013

Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Seleção de variáveis na presença de valores omissos: uma aplicação na modelação do Índice de Massa Corporal nos imigrantes africanos e brasileiros residentes em Lisboa e Setúbal

Beatriz Goulão

Trabalho de Projeto

Mestrado em Bioestatística

Orientadoras: Professora Doutora Patrícia de Zea Bermudez

Professora Doutora Valeska Andreozzi

2013

Índice

Índice de figuras

Índice de tabelas

Resumo

Os dados omissos são muito comuns em estudos clínicos e epidemiológicos. Os métodos usados por diversos programas estatísticos para tratar este tipo de problema (por exemplo, a rejeição total dos registos com observações omissas nalguma das variáveis – análise dos casos completos (CC)) nem sempre são satisfatórios. De facto, se os indivíduos com valores omissos diferirem significativamente dos com valores observados, então, não considerar os dados incompletos, poderá enviesar os resultados do estudo. Existem diversas técnicas para tratar dados omissos, nomeadamente a substituição dos valores omissos por valores considerados plausíveis, por um único valor (imputação simples) ou por vários (imputação múltipla). Esta investigação pretende avaliar o impacto de diferentes técnicas para tratamento de valores omissos na escolha de variáveis em modelos de regressão, cuja variável resposta é o índice de massa corporal (IMC). A amostra é formada por 1980 imigrantes brasileiros e africanos a viver em Portugal. Os dados foram recolhidos no âmbito do estudo de Saúde dos imigrantes, realizado em 2007. Elaboraram-se três cenários de dados omissos: 1) cenário real, com tratamento da variável com maior percentagem de dados omissos – escolaridade (6.8%); 2) simulação da existência de 20% de dados omissos na mesma variável; 3) simulação da existência de 20% de dados omissos na variável idade que está fortemente associada ao IMC. A análise CC e as técnicas de imputação conduziram a resultados semelhantes no primeiro cenário. Nos cenários 2 e 3, as técnicas de imputação revelaram-se superiores à análise CC. Os resultados deste trabalho sugerem que a existência de uma baixa percentagem de dados omissos, numa variável explicativa pouco associada com a variável resposta, parece ter poucas implicações nos resultados finais, independentemente da técnica escolhida para lidar com os dados omissos. No caso de percentagens elevadas de dados omissos, a análise CC é claramente inferior às técnicas de imputação.

Abstract

Missing data is a common problem in epidemiological and clinical studies. The methods used to handle this problem are often unsatisfactory, namely disregarding all the subjects with missing values in any of the variables used (complete case analysis). In fact, if subjects with missing data differ significantly from the remaining subjects in the sample, then using CC analysis can produce erroneous results. There are several techniques to handle missing data, including the replacement of these observations with one (simple imputation) or several (multiple imputation) plausible values. This research intends to assess the impact of different techniques to handle missing data in the selection of variables in a regression model, in which the body mass index (BMI) is the dependent variable. The sample is constituted by 1980 Brazilian and African immigrants, living in Portugal. The data were collected in 2007, as part of an Immigrants' Health study. Three scenarios of missing data were examined: 1) the real scenario, in which the variable with the highest percentage of missing data was considered (education, 6.8%); 2) simulation of 20% of missing data in the same variable; 3) simulation of 20% of missing data in the variable age, which is strongly associated with BMI. The CC analysis and the imputation techniques produced similar results in the 1st scenario. In the 2nd and 3rd scenarios, imputation techniques performed better than the CC approach. The results obtained suggest that the existence of a small percentage of missing data, in a variable poorly associated with the main outcome, seems to have little impact in final results, no matter which technique for handling missing data is used. If the percentage of missing data is high, CC analysis is clearly inferior when compared with the imputation techniques.

Agradecimentos

O meu sincero agradecimento à Professora Doutora Patrícia de Zea Bermudez e Professora Doutora Valeska Andreozzi, pelo apoio incansável, incentivo e companheirismo e pelo seu conhecimento científico que valorizou e melhorou o presente e futuros trabalhos.

À equipa responsável pelo desenvolvimento do Projeto SAIMI que, gentilmente, cedeu os dados que foram usados neste trabalho. Em particular, à Dra. Violeta Alarcão, pelos seus incentivos, esclarecimentos e enorme disponibilidade.

À minha família e amigos pelo apoio e incentivo, pela paciência e carinho que mostraram ao longo de todo o processo.

1. Introdução

O presente trabalho centra-se nas técnicas para tratar dados omissos, em estudos epidemiológicos. Em 2007, foram recolhidos dados no âmbito de um estudo de acesso aos cuidados de saúde, dos imigrantes brasileiros e africanos a viver em Portugal. Em 2013, foram publicados dados que pretendiam analisar a associação entre o índice de massa corporal (IMC) e o tempo de residência no país. A hipótese era que os imigrantes, à medida que viviam há mais anos em Portugal, aumentavam o seu IMC. Pretende-se agora estudar o impacto dos dados omissos numa variável explicativa do modelo de regressão múltiplo que analisa a associação entre anos de residência em Portugal e o IMC, assim como comparar diferentes métodos para lidar com os mesmos(1).

1.1. Dados omissos

Um questionário realizado a uma amostra da população, procura saber mais obter informações acerca de uma ou de várias características da população em causa. Fá-lo apenas a algumas unidades da mesma, aquelas que foram selecionadas para a amostra. Num estudo de recolha de dados por questionário, bem desenhado, a escolha da amostra é feita de forma cuidada, de modo a permitir estabelecer inferências para a população. No entanto, em muitos questionários, algumas das unidades contactadas não respondem, pelo menos em parte, às questões efetuadas. O problema criado pela não resposta em questionário é, obviamente, que os dados que seria suposto serem recolhidos, não existem (2).

Em todo o tipo de investigação, quer seja de carácter epidemiológico ou clínico, surgem dados omissos. Isto é particularmente verdade quando os estudos são realizados em humanos (2–4). Apesar de ser uma questão encontrada com tanta frequência na literatura, o problema dos dados omissos é frequentemente ignorado na análise estatística aplicada (5). A maioria dos *packages* estatísticos exclui os sujeitos com algum dado omisso em alguma das variáveis analisadas, tornando a análise de casos completos ou de casos disponíveis a mais comum(6). Enquanto que a maioria das técnicas estatísticas *standard* requerem uma base de dados retangular, analisar apenas os casos completos (ou seja, aqueles sem valores omissos na variável de

interesse) é ineficiente, visto que ignora informação dos respondentes que têm dados omissos nalgumas, mas não em todas, as medidas relevantes. Além disso, se a percentagem de casos com valores omissos numa ou mais variáveis for substancial, a amostra de casos completos pode não ser representativa da amostra inicial ou da população alvo do estudo (5). Esta situação pode também aumentar a probabilidade de enviesamento das mesmas, visto que os não respondentes são, com frequência frequentemente, sistematicamente diferentes dos respondentes. De particular preocupação é o facto deste enviesamento ser difícil de eliminar, visto que frequentemente habitualmente não se conhecem os motivos da não-resposta (2) e a ocorrência de dados omissos é, raramente, omissa de forma completamente aleatória (4).

Em suma, deveria ser prestada mais atenção aos dados omissos, por parte dos investigadores, tanto na fase do desenho e implementação do estudo, como posteriormente, na análise dos dados recolhidos (7).

1.2. Mecanismos de não-resposta

Rubin (1976) fez uma revisão sobre os mecanismos de não-resposta e os métodos de inferência que se utilizam na presença dos mesmos. A sua principal conclusão é que o investigador deve considerar o mecanismo que gera a não-resposta, visto que isto implica usar diferentes modelos que se adequem aos processos encontrados (8,9). Os tipos e mecanismos de não-resposta definidos por Rubin (9) são os descritos abaixo.

1.2.1. Dados omissos completamente aleatórios (MCAR)

A causa para a omissão dos dados não se encontra associada a nenhuma característica ou resposta dos sujeitos, incluindo o valor omisso, se este fosse conhecido. Impõe que a probabilidade de não-resposta seja a mesma, para diversas situações (8). Por exemplo, dados omissos num estudo de laboratório, porque o tubo de ensaio caiu ao chão, não tendo esta causa qualquer ligação com as restantes características medidas no estudo (10). Outro exemplo é a omissão de dados causada pelo término do financiamento, antes do final do estudo.

1.2.2. Dados omissos aleatórios (MAR)

Num processo MAR, a probabilidade de omissão depende somente de valores de variáveis que foram, de facto, medidas. Por exemplo, num estudo em que as mulheres

apresentem menor probabilidade de responder ao valor do seu rendimento e no qual a informação sobre o género foi recolhida (10). Com esta informação e tendo dados acerca do rendimento de algumas das mulheres na amostra, podemos fazer uma estimativa não enviesada do rendimento por sexo. Isto porque consideramos que os rendimentos que temos disponíveis para algumas das mulheres funcionam como uma amostra aleatória dos rendimentos de todas as mulheres da amostra. Outra forma de explicar os dados omissos MAR é dizer que sabendo os valores de outras variáveis disponíveis, os sujeitos com valores omissos só são aleatoriamente diferentes dos outros sujeitos. Para dados omissos MAR, a omissão de uma covariável não pode depender de fatores não observados.

1.2.3. Dados omissos não aleatórios (NMAR)

O dado omisso é não aleatório se está relacionado com fatores não observados. Neste caso, é mais provável que os elementos estejam omissos se os seus verdadeiros valores forem sistematicamente os mais elevados ou os mais baixos. Por exemplo, caso da probabilidade da indicação do rendimento ser omitida aumentar, caso sejam rendimentos muito elevados ou muito baixos.

Os dados omissos NMAR são muito difíceis de analisar não havendo, muitas vezes, forma de provar que são NMAR. A maioria dos métodos disponíveis para tratar dados omissos assume que os mesmos são MAR.

Os dados NMAR podem também ser chamados de não ignoráveis, visto que ignorá-los pode produzir enviesamentos na análise de dados (2).

1.3. Padrões de não resposta

De acordo com Rubin (2), podemos encontrar dois tipos padrões de não resposta: monotónicos ou não monotónicos. O primeiro tipo de padrão ocorre quando há dados omissos em mais do que uma variável e as colunas da base de dados podem ser "arranjadas", de modo a que $X_{i,j}$ seja observado para todos os casos em que X_j é observado (Figura 3, a)). Quando existe apenas uma variável com dados omissos, temos um caso particular do padrão monotónico. O padrão não monotónico surge quando duas variáveis nunca são observadas simultaneamente (Figura 3, b)), isto é, o

padrão é arbitrário. O segundo caso requer suposições específicas, para que possa existir um tratamento dos dados omissos (9,11).

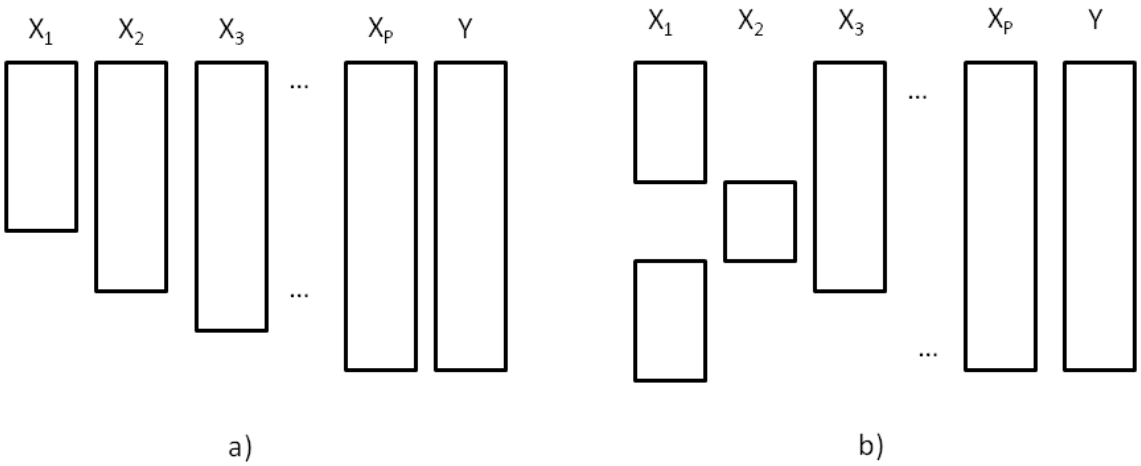


Figura 1 - Padrões de não resposta: a) padrão monotómico e b) padrão não monotómico ($X_{1,\ldots,p}$ = variáveis independentes; Y = variável dependente)

1.4.Estratégias para lidar com dados omissos

Tal como já foi referido, os métodos frequentemente usados e fáceis de aplicar para lidar com dados omissos incluem a análise de casos completos ou de casos disponíveis, ou a imputação da média. No entanto, estes métodos podem levar a análises ineficientes e, pior ainda, estimativas enviesadas das associações investigadas. Existem técnicas mais sofisticadas de lidar com dados omissos, tal como a imputação múltipla (IM), que são reconhecidamente melhores (2,9,12). Através destas técnicas de IM, o dado omisso de um certo sujeito é imputado, usando valores preditos por outras características conhecidas desse mesmo indivíduo (6).

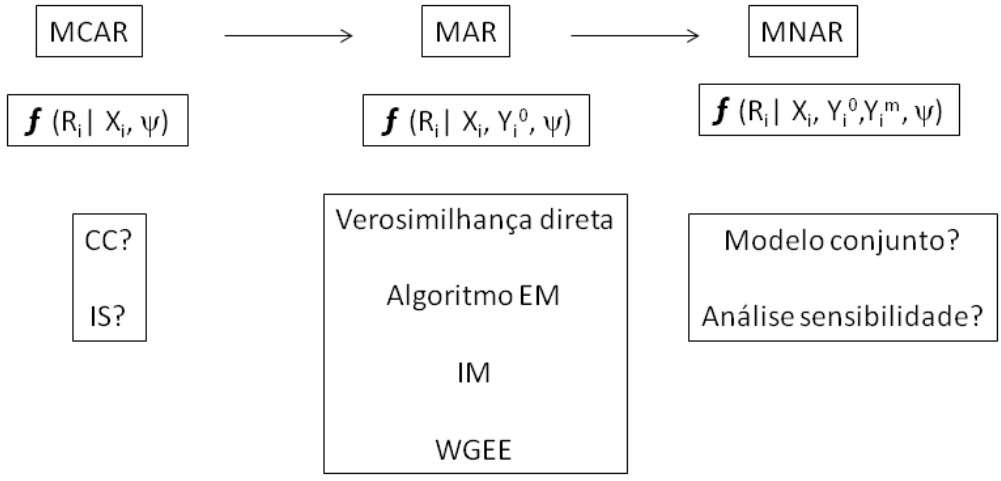
A imputação simples (IS) ou múltipla (IM) são, frequentemente, escolhidas como técnicas para tratar dados omissos. A IM é também a técnica mais recomendada no caso da recolha de dados tranversal, por questionário(4). Muitos investigadores preferem o uso da IS, em vez da IM, para evitar problemas causados por bases de dados múltiplas (13). A vantagem da IM é a incorporação da variabilidade, associada ao facto de não conhecermos realmente o dado omisso a imputar no processo de análise (2). É preciso ter em atenção o tipo de processo que gera os dados omissos que estamos a usar. O mecanismo de omissão dos dados vai sugerir as técnicas que são mais adequadas em cada caso (Figura 4). Porém, a grande dificuldade, em termos

práticos, consiste em saber qual é o tipo de mecanismo de omissão. Apenas se pode especular sobre este assunto, no sentido de investigar o mecanismo de omissão observado, mas sabe-se que é improvável, na maioria das situações, a suposição de que os dados sejam verdadeiramente MCAR.

Supondo que R_i representa uma variável indicatriz, definida por:

$$R_i = \begin{cases} 1, se\ Y_i\ \text{é}\ observado \\ 0, caso\ contrário \end{cases}$$

X_i é o conjunto de variáveis explicativas e φ um vetor de parâmetros a ser estimados. $Y_i$corresponde aos resultados obtidos do i-ésimo sujeito, sendo que Y_i^0 são os dados efetivamente observados e Y_0^m os dados que estão omissos. Pode, assim, representar-se os mecanismos de omissão, de acordo com Molenberghs. A figura 4 resume a apresentação destes mecanismos e sugere alguns métodos para lidar com cada um deles. Existem dúvidas relativamente aos melhores métodos para lidar com dados omissos, no caso do mecanismo de omissão ser MNAR. Dependendo do caso, haverá necessidade de aplicação, por exemplo, de uma técnica de imputação (simples) ou a análise de casos completos, quando o mecanismo da omissão é MCAR.



CC: análise de casos completos; IS: Imputação Simples; IM: Imputação múltipla; WGEE: Equações de estimação generelizadas e ponderadas. Adaptado de Molenberghs & Kenward, 2007.

Figura 2- Representação esquemática dos mecanismos de dados omissos, juntamente com métodos para lidar com estes e análise de sensibilidade (Adaptado de Molenberghs(3))

Quando o mecanismo de omissão é MNAR não existe nenhum método preferencial que lide com os dados omissos, de modo apropriado (4). É sempre recomendada uma

análise de sensibilidade, após o uso de um método para tratar os dados omissos MNAR, visto que os resultados podem variar muito, dependendo do modelo assumido, sendo importante experimentar diferentes modelos e ver se fornecem resultados semelhantes (14). No entanto, na investigação epidemiológica, o processo que gera os dados omissos não é frequentemente, nem MCAR nem MNAR. A omissão é tipicamente MAR, ou seja, está relacionada com outras características observadas dos sujeitos incluindo, direta ou indiretamente, a variável resposta (*outcome)* (4).

Mesmo quando o investigador não pretende tratar os dados omissos, antes de rejeitar os sujeitos em causa, deve no mínimo estudar os padrões das variáveis omissas. Isto pode ser feito através de um modelo de regressão logística ou particionamento recursivo (*recursive partioning*), de forma a predizer se as variáveis, incluindo a variável resposta, estão omissas e verificar tendências sistemáticas, em oposição a um tendência MCAR. É comum apagarem-se os indivíduos com valores omissos na variável resposta, mas em diversos modelos, pode verificar-se uma maior eficiência das estimativas dos coeficientes de regressão, utilizando as observações com dados omissos em Y (variável resposta) que não estão omissos em X (variáveis explicativas). Assim, a imputação de Y pode ter um papel importante (10).

As imputações pela média ou mediana, por regressão ou regressão estocástica são todas baseadas na seguinte equação:

$$v_i^* = a + X_i^T b + e_i^*, i = 1, \dots, N_m \text{ [1]}$$

onde v_i^*é o valor imputado para uma resposta omissa na variável $v$para o caso i da amostra, X_i é a K-coluna do vetor de observações dos K preditores de regressão para o caso $i$do modelo de imputação, bé um vetor coluna de ordem K dos coeficientes de regressão estimados que correspondem às variáveis X, e_i^*é o resíduo da regressão de $v$em X e N_m é o número de casos da amostra para os quais é necessário aplicar-se a imputação.

As estratégias para lidar com dados omissos podem incluir a imputação. A imputação é um método estatístico para o tratamento da não resposta, sendo os valores ausentes

substituídos pelas suas estimativas (15). Existem diversas técnicas de imputação que se dividem em vários grupos, como apresentado de seguida.

- **Métodos dedutivos**: o valor imputado é deduzido a partir de informação conhecida, nomeadamente em inquéritos anteriores que utilizam as mesmas questões e amostra.
- **Métodos determínisticos**: baseia-se nos dados de todos os respondentes. Para unidades com as mesmas características produz sempre o mesmo valor imputado. Englobam-se neste método a imputação pela média ou mediana e a imputação por métodos regressivos.
- **Métodos estocásticos**: produzem imputações diferentes sobre a mesma unidade de não resposta parcial. Os mais usuais são os métodos de Hot-Deck (*predictive mean matching* e índice de propensão, por exemplo), a imputação por associação flexível e os métodos regressivos com efeito aleatório.

O presente trabalho centra-se na aplicação de técnicas para lidar com dados omissos, do tipo MAR e MCAR. As técnicas para lidar com dados omissos do tipo MNAR não são exploradas. Adaptar as escolhas destas técnicas ao tipo de dados omissos de interesse (variável contínua, categórica ou dicotómica) é essencial. Neste caso, foram selecionadas técnicas de IS e IM, de modo a lidar com dados omissos em variáveis contínuas.

1.5. Motivação: a saúde dos imigrantes

1.5.1. A variável resposta: O índice de massa corporal

O IMC é um indicador de risco cardiovascular, amplamente usado pela sua simplicidade. O seu cálculo é feito a partir de dados de peso e altura, correspondente à razão peso(kg)/altura^2(metros^2). Esta variável é, frequentemente, classificada de acordo com os critérios estipulados pela Organização Mundial de Saúde (< 18,5 kg/m^2 – Magreza; 18,5 a 24,9 kg/m^2 – Eutrofia ou Normoponderal; 25 – 29,9 kg/m^2 – Pré-obesidade; ≥ 30 kg/m^2 – Obesidade)(16). De acordo com o mesmo critério, o excesso de peso é a junção das categorias pré-obesidade e obesidade, ou seja, IMC ≥ 25 kg/m^2.

Ter excesso de peso está associado a um risco aumentado para desenvolver diversas patologias, nomeadamente diabetes, hipertensão arterial, doenças cardiovasculares e diferentes tipos de cancro (17).

A prevalência de excesso de peso tem vindo a aumentar, nas últimas décadas (18). Embora saibamos que os maus hábitos alimentares e a elevada prevalência de sedentarismo estão associados ao aumento do excesso de peso, há uma grande dificuldade em travar o processo.

Compreender as variáveis associadas ao IMC é de extrema importância, para que se possam delinear estratégias eficazes na prevenção do excesso de peso.

1.6. Os imigrantes brasileiros e africanos a viver em Portugal

A população estrangeira a residir, de forma legal, em Portugal tem vindo a aumentar nos últimos anos. De facto, em 2005, havia 430.747 imigrantes (19), sendo 457 306 em 2010 (dados publicados pelo Serviço de Estrangeiros e Fronteiras) (20).

Pode constatar-se, através da leitura da Figura 1, que cerca de metade dos imigrantes a residir no país são ou provenientes de países de língua oficial portuguesa (PALOP) (21%) ou do Brasil (25.5%).

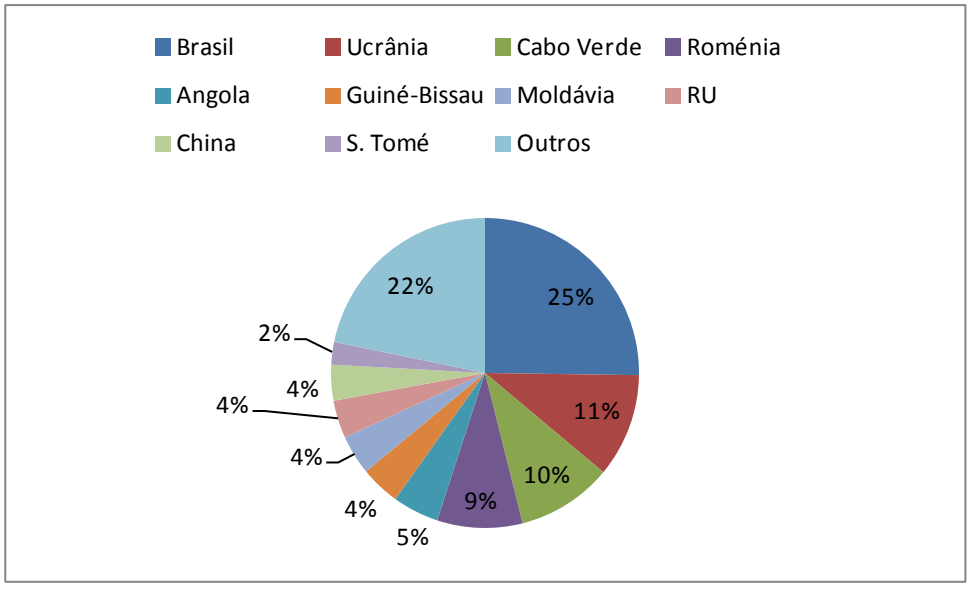


Figura 3- Prevalência de imigrantes a residir em Portugal, por nacionalidade (20)

Os imigrantes são um grupo de risco no desenvolvimento de excesso de peso. Pertencem, frequentemente, a níveis socioeconómicos mais baixos do que os nativos,

têm acessos aos cuidados de saúde mais dificultados, havendo também uma maior sensibilidade étnica para o desenvolvimento de doenças cardiovasculares, no geral (21). Apesar disto, a investigação nesta área e, em particular, em Portugal é ainda incipiente e há ainda muito por conheceracerca dos níveis de saúde dos imigrantes.

1.7. A aculturação alimentar e o impacto na saúde dos imigrantes

O termo aculturação é frequentemente usado para denominar um processo pelo qual um grupo étnico, usualmente uma minoria, adota os padrões (por exemplo, crenças, linguagem e (ou) dieta) do grupo de acolhimento/dominante. A aculturação alimentar refere-se ao processo que ocorre quando membros do grupo minoritário adotam os padrões alimentares/escolhas alimentares do país de acolhimento (22).

A migração populacional implica, frequentemente, alterações no tipo e qualidade dos alimentos que os migrantes consomem e a sua forma de preparação. Os tipos de alimentos que mudam mais rapidamente são "alimentos acessórios", como *snacks* e doces, enquanto os alimentos que permanecem inalterados por um período mais longo de tempo são "alimentos base", como por exemplo o arroz e o milho (23). Muitas vezes, este processo pode ter contornos negativos. Sabe-se, por exemplo, que a aculturação à dieta americana resulta num aumento na ingestão de gordura, sal, carne, leite e açúcar e numa diminuição do aporte de hidratos de carbono complexos, fibra e muitas vitaminas e minerais (24). Os adolescentes africanos, a viver em Portugal, recorrem ao consumo de refrigerantes e *fast-food*, principalmente para estar com os seus pares (25).

Estas alterações longitudinais nas escolhas alimentares são determinadas por fatores como a disponibilidade e o preço dos alimentos no país de acolhimento (26,27), o rendimento(26), a idade dos imigrantes aquando da imigração (28), crenças alimentares (26) e o *stress* aculturativo(29) (fenómeno caracterizado por sentimentos de solidão e isolamento e que pode resultar em comportamentos alimentares pouco saudáveis e na diminuição da atividade física).

Uma maior aculturação e aculturação alimentar estão associadas a um maior risco de vir a ter excesso de peso nos imigrantes a viver em países como os EUA, Canadá e Austrália (30). Na Europa, também existem alguns dados que apontam para uma

associação positiva entre estas variáveis (31,32). Em Portugal, detectou-se uma associação positiva e significativa entre o tempo de residência no país e o IMC, nos imigrantes brasileiros e africanos (1).

As hipóteses propostas, no sentido de compreender a associação entre tempo de residência no país alóctone e o IMC, são complexas e envolvem diversas variáveis. A Figura 2 mostra algumas das variáveis que tendem a ser incluídas neste processo, como moderadoras do efeito que a aculturação (medida, neste caso, pelos anos de residência no país de acolhimento) provoca.

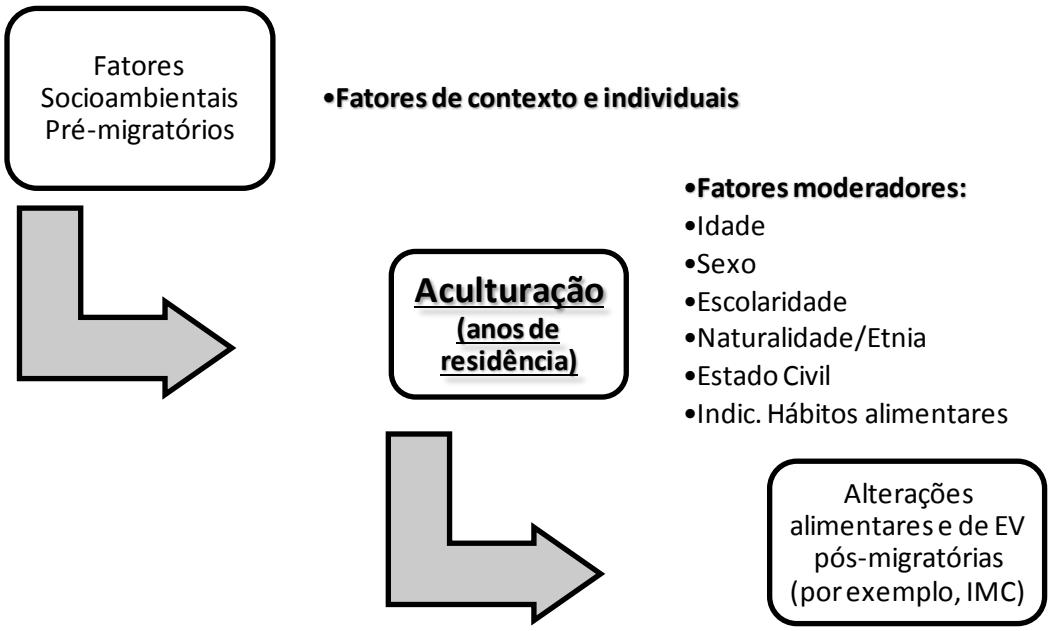


Figura 4 - Modelo explicativo do impacto da aculturação na saúde dos imigrantes e seus mediadores (EV: estilos de vida)

É essencial compreender o contexto pré-migratório dos imigrantes. O país de origem tem importância no impacto da aculturação, na saúde dos mesmos. Isto acontece, dado que existem fatores genéticos e culturais, de extrema importância, neste contexto (33). Fatores individuais, como o motivo que levam o imigrante a imigrar, são também relevantes. O processo de aculturação é, frequentemente medido pelo tempo de residência num país de acolhimento. Esta variável *proxy* parece ser essencial para compreender o processo (30). Fatores como a idade, o sexo, a escolaridade, a naturalidade, o estado civil e os hábitos alimentares podem e devem ser tidos em conta, num modelo múltiplo, que tente compreender esta associação, como é explicado em Goulão, 2013 (1). Este processo leva, inevitavelmente a alterações

alimentares e de estilo de vida que terão impacto nos níveis de saúde dos imigrantes (Figura 2).

2. Objetivos

Objetivo geral

2.1. Comparar diferentes abordagens para o tratamento de valores omissos na seleção de variáveis explicativas associadas ao IMC dos imigrantes africanos e brasileiros residentes em Lisboa e Setúbal

Objetivos específicos

2.1.1. Comparar a análise de casos completos, as técnicas de imputação simples por mediana, *predictive mean matching* e índice de propensão e as técnicas de imputação múltipla por *predictive mean matching* e regressão linear não Bayesiana para lidar com dados omissos na variável escolaridade, que no estudo mencionado apresenta 6.8% de valores omissos, considerando um mecanismo MAR.

2.1.2. Comparar a análise de casos completos, as técnicas de imputação simples por mediana, *predictive mean matching* e índice de propensão e as técnicas de imputação múltipla por *predictive mean matching* para lidar com dados omissos na variável escolaridade, parcialmente retirados de forma aleatória, com 20% de omissão, considerando um mecanismo MAR.

2.1.3. Comparar a análise de casos completos, as técnicas de imputação simples por mediana, *predictive mean matching* e índice de propensão e as técnicas de imputação múltipla por *predictive mean matching* para lidar com dados omissos na variável idade (fortemente associada ao IMC), totalmente retirados de forma aleatória, com 20% de omissão, considerando um mecanismo MCAR.

3. Inquérito da Saúde dos Imigrantes

3.1. Projeto SAIMI

O projeto "Acesso aos Cuidados de Saúde e Nível de Saúde das Comunidades Imigrantes Africana e Brasileira em Portugal" (SAIMI) foi realizado no Instituto de Medicina Preventiva, da Faculdade de Medicina da Universidade de Lisboa, e teve como objetivos a caracterização do estado de saúde das comunidades imigrantes entre si e a comparação do estado de saúde desta população com o da população Portuguesa em geral. A recolha de dados foi feita através de um questionário adaptado do que foi utilizado no 4º Inquérito Nacional de Saúde (INS). Pretendeu também caracterizar o acesso dos imigrantes aos cuidados de saúde e a prestação efetiva de cuidados a estas populações. Os investigadores responsáveis pela projeto foram o Dr. Mário Carreira, o Dr. Rui Portugal e a Dra. Violeta Alarcão, da Faculdade de Medicina da Universidade de Lisboa.

Foi realizada uma análise de dados secundária ao Projeto SAIMI, no âmbito da dissertação em Doenças Metabólicas e Comportamento Alimentar, denominado "Excesso de peso nos imigrantes brasileiros africanos: prevalência e associação com o tempo de residência em Portugal" (1) com o propósito de caracterizar a prevalência de excesso de peso nos imigrantes brasileiros e africanos incluídos no estudo e analisar a associação entre o excesso de peso e índice de massa corporal e com os anos de residência no país de acolhimento, (Portugal).

3.2. Recolha de dados

Numa primeira fase, foram pré-selecionados dois distritos (Lisboa e Setúbal) , por serem os que apresentam maior proporção de imigrantes em Portugal Continental, de acordo com os Censos 2001. Em seguida e, de acordo com o mesmo critério, pré-seleccionaram-se treze concelhosnessas dois distritos. Estes concelhos foram: Sintra, Lisboa, Loures, Amadora, Cascais, Odivelas, Oeiras, Vila Franca de Xira (distrito de Lisboa); Seixal, Almada, Setúbal, Moita, Barreiro (distrito de Setúbal). Este processo permitiu incluir, na amostra, 98,2% dos imigrantes que vivem no distrito de Lisboa e 90,6% dos imigrantes que vivem no distrito de Setúbal. A partir da seleção dos concelhos, implementou-se uma amostragem aleatória espacial por *clusters*, num passo apenas, em que a seleção dos *clusters* obedeceu a uma amostragem espacial

aleatória simples. Este processo foi feito através do programa *ArcMap* que seleciona polígonos com dimensão 50 x 50 m^2 (constituindo cada um destes, a unidade amostral primária, ou seja, os *clusters*). O programa abrangeu todo o território selecionado e não apenas as áreas ocupadas por habitação. Assim, foi necessário verificar, por interpretação visual dos 20 *clusters* selecionados, se os *clusters* continham pelo menos uma habitação (critério de inclusão) ou não (critério de exclusão). Os processos de amostragem e validação foram repetidos várias vezes, até se obterem 20 *clusters* válidos. Após obtenção dos mesmos, iniciou-se a preparação de um relatório destinado aos entrevistadores que continha imagens dos locais a visitar, no terreno, e o percurso mais eficaz para os alcançar. As ferramentas de *routing* utilizadas para localizar os *clusters* foram o *Multimap* ou *Google Earth*. No terreno, foram também usados GPS (*Global Positioning System*), de modo a facilitar a orientação dos entrevistadores. As equipas de entrevistadores, compostas por duas pessoas, preferencialmente oriundas das comunidades em estudo, visitaram todos os domicílios incluídos em cada *cluster* (unidade amostral secundária), convidando as pessoas elegíveis a responder ao questionário. Caso a entrevista não pudesse ocorrer no momento da visita, os entrevistadores voltavam às habitações até conseguirem entrevistar todos os imigrantes elegíveis.

3.3. Amostra

Os imigrantes selecionados para efeitos do presente estudo são os mesmos usados na presente análise e foram selecionados de acordo com os seguintes critérios de inclusão:

Foram incluídos imigrantes residentes em Lisboa e Setúbal (à data da recolha dos dados) e que tenham: nascido num país PALOP e que tenham vindo para Portugal após 1980; ounascido no Brasil e que se considerem na situação de imigração desde 1995. Ou seja, para efeitos desta análise considerámos apenas os imigrantes de 1ª geração, nascidos em país estrangeiro. Os outros dois critérios de inclusão usados foram a idade (entre os 18 e 64 anos) e terem respondido às questões acerca do peso e altura, de modo a permitir o cálculo do IMC. A amostra final ficou assim constituída por **1980 indíviduos**.

3.4. Instrumento de recolha de dados

O questionário usado foi adaptado do 4º Inquérito Nacional de Saúde, realizado em 2005/06, e possuia os seguintes domínios:

- *Caracterização sociodemográfica;*
- *Trajetória imigratória:*
- *Acessibilidade aos cuidados de saúde;*
- *Saúde reprodutiva;*
- *Informações gerais de saúde (autoavaliação do estado de saúde, dados antropométricos e outros);*
- *Doenças crónicas;*
- *Despesas de saúde e rendimentos;*
- *Consumo de tabaco;*
- *Consumo de alimentos e bebidas;*
- *Saúde infantil;*
- *Atividade física;*
- *Saúde mental e bem-estar geral;*
- *Saúde Oral.*

Para efeitos da presente análise foram selecionados os domínios caracterização sociodemográfica (idade, sexo, escolaridade e estado civil), trajetória imigratória (anos de residência em Portugal e naturalidade), informações gerais de saúde (peso e altura autorrelatados) e consumo de alimentos e bebidas (número de refeições principais e intermédias consumidas diariamente). As páginas do questionário que dizem respeito a estas questões encontram-se no Anexo 9.2.

3.5. Breve descrição das variáveis e caracterização da amostra

As variáveis usadas para a presente análise encontram-se na Tabela 1. Uma breve caracterização dos dados é apresentada na Tabela 2. Nesta tabela, salienta-se particularmente a última coluna que apresenta a percentagem de valores omissos observada em cada variável.

Tabela 1 – Variáveis do projeto SAIMI usadas no presente estudo

Nome	Nome na base de dados	Opções de resposta
IMC (calculado com base no peso e altura autorelatados)	imc	Em quilogramas por metro quadrado, com uma casa decimal
Sexo	sx	0 = Feminino; 1 = Masculino
Idade	idade	Dos 18 aos 64 anos
Escolaridade	escol	Anos de escolaridade
Estado civil	estcivil	0 = solteiro; 1 = casado; 2 = divorciado; 3 = viúvo
Estado civil recategorizado	estcivilr	0 = solteiro; 1 = casado; 2 = outro
Origem	origem	0 = africanos; 1 = brasileiros
Anos de residência em Portugal	anos	≥ 0 anos
Número de refeições principais	refeicoes	0 = Uma; 1 = Duas; 2 = Três
Número de refeições principais recategorizado	refeicoescat	0 = Uma ou duas; 1 = Três
Número de refeições intermédias	*snack*	≥ 0 refeições
Número de refeições intermédias categorizadas	*snackr*	0 = 0; 1 = 1; 2 = 2; 3 = Três ou mais

Tabela 2 – Caracterização da amostra

Variáveis	N = 1980	% Dados omissos
Sexo	Feminino: 1058 (53.4%) Masculino: 922 (46.6%)	0
IMC	25.07 ± 4.46 kg/m2	0
Idade	35.1 ± 10.95 anos	0
Estado civil	Solteiro: 744 (37.6%) Casado: 1100 (55.6%) Outro: 130 (6.6%)	0.3
Escolaridade	9.21 ± 3.53 anos	6.8
Origem	Africanos: 1080 (54.6%) Brasileiros: 705 (35.6%)	0
Anos residência em Portugal	9.84 ± 8.14 anos	1.2
Nº refeições principais	<3: 537 (27.1 %) 3: 1426 (72.0 %)	0.9
Nº refeições intermédias	0: 560 (28.2%) 1: 737 (37.2%) 2: 404 (20.4%) 3 ou mais: 229 (11.6%)	2.6

4. Métodos

Para efeitos da análise de dados omissos realizou-se um estudo exploratório dos mesmos que permitisse compreender os principais padrões de omissão e variáveis associadas aos mesmos. Para isso, recorreu-se a métodos estatísticos como a análise de clusters e a regressão logística com resposta binária (ser ou não ser omisso em dada variável), tal como preconizado por Harrell (10). Usalmente, as análises univariadas não fornecem a totalidade da informação acerca dos dados omissos, sendo importante perceber a percentagem de dados omissos nas diferentes associações entre variáveis (34). Embora seja impossível estabelecer se a omissão de uma dada variável é MAR ou MNAR, pode especular-se sobre o assunto, de forma a clarificarmelhor os motivos que ditam a omissão(9,12). Um investigador não deve optar por uma análise de casos completos, sem antes "explorar" os seus dados omissos, de modo a evitar enviesamentos graves nos resultados e na sua interpretação (10). Na segunda fase deste presente estudo, optou-se pelo uso de técnicas de imputação para lidar com os dados omissos, em oposição a outras técnicas que podem ser usadas para o mesmo efeito, nomeadamente métodos com estimação baseados na máxima verosimilhança.

4.1. Escolha das variáveis com valores omissos a serem analisadas

Partindo da amostra de 1980 indivíduos do projecto SAIMI, para efeitos do presente estudo, optou-se por realizar três diferentes cenários, relativamente aos dados omissos:

1º cenário: cenário real

Tratamento da variável com maior percentagem de dados omissos – escolaridade (6.8%) Os casos das restantes variáveis explicativas com valores omissos foram eliminados por serem em número muito reduzido (o que levou a uma percentagem de dados omissos na variável idade ligeiramente inferior à inicial – 6.3%).

2º cenário:

Simulação da existência de 20% de dados omissos na variável escolaridade.

3º cenário:

Simulação da existência de 20% de dados omissos na variável idade. A seleção desta variável foi motivada pelo facto de ser a que está mais fortemente associada com a variável resposta.

As simulações foram realizadas de modo a garantir a aleatoriedade dos dados omissos fabricados. No primeiro caso, retirou-se a percentagem de dados necessária para completar os 20% de dados omissos. Ou seja, neste caso nem todos os dados omissos foram escolhidos de modo aleatório, tendo-se mantido os que já existiam. Na segunda simulação, a totalidade dos casos assumidos como omissos foram escolhidos de forma aleatória.

4.2. Análise dos casos completos

Uma análise de casos completos (CC) inclui apenas os casos para os quais todas as variáveis foram recolhidas. Este método apresenta vantagens que se prendem com a sua simplicidade de aplicação e com a possibilidade de se usarem ferramentas estatísticas usuais, já que a estrutura dos dados é a esperada. No entanto, apresenta diversos constrangimentos. Há, frequentemente, uma perda significativa de informação, fazendo com que as estimativas dos parâmetros dos modelos não sejam eficientes. Além disso, os resultados poderão ter um viés grave se se tratarem de dados omissos MAR, ao contrário de dados MCAR. Além do possível viés encontrado nos resultados, podem levantar-se dois problemas principais com o uso de uma análise de casos completos:

1. Se existirem muitas variáveis a incluir no modelo, podemos ter um número reduzido de casos completos para efetuar as análises necessárias (35).

2. A redução do número de sujeitos na amostra leva a que os desvios-padrão amostrais aumentem, os intervalos de confiança dos parâmetros apresentem uma amplitude elevada e a potência dos testes de associação e de ajustamento diminua (10,36).

4.3. Imputação Simples

4.3.1. Imputação por substituição não condicional da mediana

O método de imputação pela mediana ou média é um método não condicional, dado que se substitui o valor omisso pela mediana ou média dos valores observados da mesma variável nos restantes sujeitos. Ou seja, o termo não condicional refere-se ao

facto de que o investigador não usa informação acerca do sujeito para o qual a imputação é gerada (3).

No caso da imputação pela média ou mediana, os coeficientes b e os resíduos e_i^* na equação [1] são estipulados como zero e os valores omissos são substituídospela média ou mediana de v das observações na amostra.

Este tipo de imputação, tal como todas as técnicas de imputação simples, assume que o dado imputado é o verdadeiro, não tendo em conta a variabilidade associada à imputação. Por este motivo, os resultados de análises quantitativas, como o cálculo de correlações, obtidos a partir desta imputação podem estar severamente enviesados (2). Se estivermos perante uma variável contínua ou binária X que não está associada aos restantes X, a média ou mediana podem ser usadas como substitutas dos valores omissos, sem grande perda de eficiência, embora os coeficientes de regressão possam estar enviesados, por subestimação, já que o Y não é utilizado na imputação. Quando a variável de interesse está associada a outros X, é muito mais eficiente usar um modelo preditivo individual para X, baseado noutras variáveis (10). A vantagem deste método é a facilidade de implementação e compreensão. No entanto, pode distorcer gravemente a distribuição da variável em causa, levando a complicações no sumário das medidas, incluindo, notavelmente, subestimações do erro padrão. Além disso, a imputação por substuição pode distorcer associações entre variáveis, tendo a tendência de "puxar" as estimativas de correlação para zero (35).

4.3.2. Imputação por *Hot-Deck*

O método *Hot-Deck* apresenta diferentes significados em diferentes fontes literárias (15). A imputação por *Hot-Deck* envolve a substituição dos valores omissos de uma ou mais variáveis de um não-respondente com valores observados de um respondente ou doador que é similar ao não-respondente, no que diz respeito a características observadas em ambos os casos (37). O termo *Hot-Deck* é usado em contraste com os métodos C*old-Deck*, nos quais se usam imputações de uma base de dados prévia. Um exemplo simplificado da imputação *Hot-Deck* encontra-se na Figura 5. Selecionam-se os sujeitos com valor observado na variável de interesse, escolaridade, com perfis mais similares aos sujeitos com valor omisso nessa variável. Neste caso, foram usadas as variáveis sexo, grupo etário, estado civil e origem para tomar esta decisão. No caso da

imputação por *Hot-Deck* mais simples, as variáveis a serem tidas em conta para o emparelhamento, devem ser todas do tipo categórico. Assim, o sujeito 3 passaria a ter 11 anos de escolaridade (a partir do sujeito 5, com que partilha o grupo etário, estado civil e origem), o sujeito de 7 passaria a ter 7 anos de escolaridade (a partir do sujeito 8, com quem partilha o sexo, grupo etário e estado civil) e, por fim, o sujeito 10 teria 9 anos de escolaridade (a partir do sujeito 9, com quem partilha o sexo, grupo etário e origem).

ID	Sexo	Grupo etário	Estado Civil	Origem	Escolaridade
1	F	2	S	A	13
2	F	3	C	B	12
3	F	2	C	B	-
4	F	3	C	A	4
5	M	2	C	B	11
6	F	3	S	B	12
7	M	1	S	B	-
8	M	1	S	A	7
9	F	4	V	A	9
10	F	4	D	A	-

Legenda: ID – Identificação na base de dados; Sexo: F – Feminino, M – Masculino; Grupo etário: 1: 18 – 25; 2: 25 – 34; 3: 35 – 44; 4: 45 – 54; 5: 55 – 64; Estado civil: S – Solteiro; C – Casado; V – Viúvo; D – Divorciado; Origem: A – Africana; B - Brasileira

Figura 5 - Exemplo da imputação *Hot-Deck* simplificada

A imputação deste tipo pode distinguir-se em dois grupos: os métodos de *Hot-Deck* aleatórios, nos quais o dador é selecionado aleatoriamente de um grupo de potenciais dadores, que se pode chamar uma *pool* de dadores; os métodos *Hot-Deck* determínisticos que identificam um dador e imputam o valor desse caso, usalmente a partir do vizinho mais próximo, escolhido com base num cálculo métrico (37).

As vantagens dos métodos de *Hot-Deck*, em estudos tranversais, são diversas. Como em todos os métodos de imputação, obtém-se uma base de dados retangular, permitindo a utilização de análises estatísticas com métodos tradicionais. Não se apoia

na modelação da variável a ser imputada e, por isso, é potencialmente menos sensível a erros na especificação do modelo, em comparação com um modelo paramétrico, tal como a imputação por regressão. Apesar disto, é importante ter em conta que o *Hot-Deck* tem pressupostos implícitos, como a escolha da medida para emparelhar dados e recetores de dados, ou as variáveis a incluir no modelo. Outro atrativo desta técnica é que apenas dados plausíveis podem ser imputados, já que os valores são selecionados a partir de valores observados, de uma *pool* de dadores. Pode existir um ganho na eficiência, em comparação com a análise de CC, visto que a informação dos casos incompletos é incluída. Existe também uma redução no enviesamento, por não resposta, visto que há uma associação entre as variáveis que definem uma classe de imputação, a propensão para responder e a variável a ser imputada (37).

Este tipo de imputação pode ser usada em combinação com a regressão, definindo a "semelhança" como a proximidade ao valor predito num modelo de regressão. Um exemplo deste tipo de imputação é o caso em que é necessário encontrar-se os fatores de risco para novos casos de VIH (vírus da imunodeficiência humana). Os fatores de risco são obtidos a partir da leitura da ficha médica de cada sujeito, mas para muitos existe falta de informação. Para cada um destes casos "não resolvidos", propôs-se uma imputação aleatória dos fatores de risco dos cinco casos "resolvidos" mais próximos. Estes casos foram definidos com base num *score* que penaliza diferenças no sexo, idade, a clínica onde os testes foram feitos, entre os fatores disponíveis em todos ou na maioria dos casos (35).

Quando usamos este tipo de técnicas, como é o caso da técnica do "vizinho mais próximo", os valores omissos são imputados sob o pressuposto que os casos com variáveis independentes semelhantes têm respostas similares. Quando estamos perante uma situação de múltiplas variáveis independentes é mais complicado encontrar observações com os mesmos valores nas variáveis independentes X. Mesmo no caso simples em que todas as covariáveis são binárias, existirão $2p$ valores possíveis para X, onde p é a dimensão de X. Isto torna difícil encontrar pares homogéneos em X (13).

Neste trabalho, o método do *Hot-Deck* é aplicado, através de duas abordagens: o *Predictive Mean Matching* e o índice de propensão, que serão desenvolvidas adiante.

4.3.3. Imputação através da aplicação do *predictive mean matching*

O *Predictive Mean Matching* (PMM), enquanto técnica de imputação simples, foi desenvolvido por Little (1988) (38). Baseia-se nos seguintes passos:

1. Estimação de um modelo de regressão, sendo a variável de interesse (a imputar) a variável resposta e as restantes variáveis recolhidas, as explicativas.

2. Estimação do valor da variável de interesse para os sujeitos com o dado omisso.

3. Emparelhamento do valor da variável de interesse predito, para os sujeitos com o dado omisso, com o valor ajustado mais próximo (feito a partir do cálculo da distância euclidiana). É imputado o valor observado correspondente ao valor ajustado mais próximo. Caso exista mais do que um valor ajustado com distância igual à distância mínima encontrada, então escolhe-se aleatoriamente o valor a imputar, de entre os que sofreram empate.

Para este efeito, foi usada uma função no programa estatístico R criada por Andreozzi (39).

No caso da imputação por regressão estocástica, todos os termos em [1] são diferentes de zero, já que cada valor omisso é substituído pelo valor predito a partir da regressão mais o resíduo. Este tipo de regressão permite estimar a média e variância da variável imputada e, como resultado, os erros padrão não são enviesados. A adição de um resíduo aleatório ao valor imputado aumenta a variância da imputação (variância associada à incerteza sobre que valores imputar) e reduz a precisão dos dados imputados. O resíduo na equação 1 reduz-se a:

$$\varepsilon_i^* = v_i^* - (a + X_i b)$$

em que v_i^* é o valor de v obtido a partir do dador. Pelo facto dos resíduos não serem "forçados" a ter a mesma média e variância entre níveis da distribuição preditiva, escolher um resultado de um dador próximo oferece proteção contra uma incorreta suposição de homocedasticidade e uma incorreta especificação do modelo (5).

As técnicas de regressão estocástica são superiores à imputação pela mediana e aos métodos de regressão determinísticos, no que diz respeito à estimação de erros padrão. Comparativamente com outras técnicas estocásticas, o PMM é facilmente operacionalizável e, visto que os valores omissos são imputados a partir de valores observados, é apropriada para a imputação de valores discretos ou contínuos. Embora esta técnica seja melhor do que outras na estimação de variâncias e erros padrão, continua a verificar-se a atenuação de associações entre a variável de interesse, e outras variáveis não incluídas na equação de imputação. Esta atenuação pode enviesar os efeitos da regressão não só no rendimento mas também nas outras variáveis independentes de interesse e o viés nas outras variáveis de interesse aumenta, à medida que as correlações com o rendimento aumentam. Um segundo problema – partilhado com as restantes técnicas de imputação simples – é a negligência da incerteza inerente acerca de que valores imputar, não levando em consideração a variância da imputação. Os erros padrão estão, por isso, enviesados em direção a 0 e os erros Tipo I inflacionados (5).

No que toca à eficácia da aplicação do PMM, a capacidade de predição do modelo é de grande importância, melhorando a precisão das estimativas dos coeficientes de regressão, nos modelos com dados imputados, tanto ou mais do que a imputação múltipla (5).

4.3.4. Imputação *Hot-Deck* com índice de propensão

Outra abordagem *Hot-Deck* é o índice de propensão (IP), em que se pretende imputar um valor do sujeito com valor observado mais semelhante àquele cujo valor está omisso. Em geral, definir a semelhança entre sujeitos pode ser um procedimento complexo. De um ponto de vista metodológico, é frequentemente necessário uma abordagem multivariada. É comum existir este problema em estudos observacionais, em que os doentes não são selecionados aleatoriamente para um tratamento. Neste caso, é possível que variáveis explicativas e potenciais confundimentos não estejam igualmente distribuídos, produzindo estimativas enviesadas. Uma abordagem que pretende evitar este enviesamento é o IP (40). O IP pode ser definido como a probabilidade de um sujeito ser exposto a um determinado tratamento, dado uma

estrutura de variáveis explicativas, X. O IP é tipicamente estimado via regressão logística múltipla (41).

O IP pode ser usado, por exemplo, para ajustar o efeito de uma potencial variável de confundimento numa análise de regressão ou para criar um emparelhamento entre sujeitos expostos e não expostos. Se o status de exposição for definido como um indicador da presença de valores omissos numa variável da análise, então o IP pode ser usado para avaliar a semelhança entre doentes com ou sem dados omissos.

O método do "propensity matching" ou IP, como técnica para lidar com dados omissos,foi proposto por Mittinty et al. (13), e foi desenvolvido por Rosenbaum e Rubin (1983) (42), num contexto diferente.

Assim sendo, o IP é definido da seguinte forma (13):

$$\pi(\mathrm{R}) = \frac{\mathrm{e}^{\beta \mathrm{X}^{T}}}{1 + \mathrm{e}^{\beta \mathrm{X}^{T}}}$$

R representa uma variável indicatriz que assume o valor 0 se a variável tem valor observado e 1 se a variável tem valor omisso. β representa o vetor de coeficientes, $\beta = (\beta_0, \beta_1, \dots, \beta_P)$, das variáveis explicativas, X^T.O IP sumariza a informação extensiva acerca de uma matriz de covariáveis numa única variável. A validade das semelhanças estabelecidas pelo IP depende da qualidade e quantidade de variáveis explicativas usadas (40).

Para se proceder ao emparelhamento dos sujeitos, pode-se usar o IP para criar classes de imputação (14,43) ou recorrer à abordagem do vizinho mais próximo (13,40).

Sendo Z a variável de interesse a imputar, a imputação, a partir do IP, envolve os seguintes passos (14):

1. Calcular um modelo múltiplo de regressão logística, cuja variável resposta é se Z está ou não omisso. As variáveis independentes selecionadas para o modelo são opção do investigador.
2. Usar o modelo logístico estimado, na alínea anterior, para calcular a probabilidade predita da variável Z estar omissa (denominada de IP).

3. Dividir as observações por IP, fazendo grupos de acordo com os quantis;
4. Em cada quantil, existem r casos com a variável Z observada e m casos com a variável Z omissa. De entre os r casos observados, aplicar uma técnica de reamostragem aleatória com reposição. Para cada caso omisso, retirar aleatoriamente um valor (com reposição) da amostra aleatória de r casos e usar o valor observado de Z como o valor imputado. Esta técnica é chamada de *bootstrap* Bayesiano.

Este algoritmo foi programado em R, para efeitos de aplicação nesta tese, por Andreozzi (39).

É importante ter em conta que este método tem limitações, visto que usa apenas as variáveis associadas com a omissão da variável Z. Não usa correlações entre as variáveis explicativas. É eficiente para fazer inferências acerca da distribuição das variáveis imputadas, tal como análise univariada, mas pode não ser apropriado para análises que envolvam associações entre variáveis, tais como modelos de regressão (14).

4.4. Imputação múltipla

A imputação múltipla (IM) foi introduzida há cerca de trinta anos, por Rubin (1978) (9). Os seus domínios de aplicação são, atualmente, variados passando por estudos observacionais de saúde pública e incluindo estudos clínicos (3). A IM começou a ser discutida nos anos 70, por Rubin, e tem ganho popularidade em publicações científicas, em particular, desde os anos 90. A sua utilização tem sido crescente e pode constatar-se que num ritmo mais acelerado, em publicações que a aplicam, como método estatístico (abstracts), sem se centrar na técnica, em comparação com as publicações cujo enfoque principal é a técnica de IM (Figura 6).

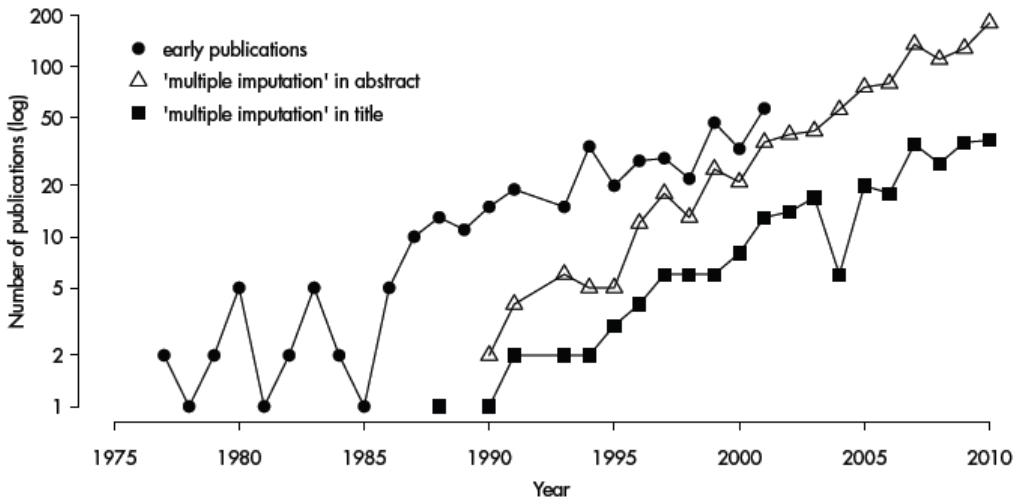


Figura 6- O aumento da popularidade da IM(44)

Em muitos aspetos, a IM mantém as vantagens da IS (nomeadamente, a possibilidade de se aplicarem métodos estatísticos para dados completos e a capacidade de incorporar informação do investigador), apresentando o benefício de corrigir as suas maiores falhas. A IM permite ao investigador usar o seu conhecimento para refletir a incerteza acerca dos valores a imputar. Esta incerteza resulta do facto da recolha de dados ser realizada numa amostra e não na população (medida convencional de variabilidade), da variância causada pela existência de valores omissos nesta amostra e da variância da simulação, causada pelo facto dos seus cálculos serem também baseados num número finito M de valores (2,44).

O primeiro estadio da IM consiste em criar múltiplas cópias de uma base de dados, com os valores omissos substituídos pelos valores imputados. Estes são selecionados a partir de uma amostra de valores com base na sua distribuição predita, tendo em conta os valores observados – assim, a IM é baseada numa abordagem Bayesiana. O procedimento da imputação tem de ter em conta a incerteza na predição de valores omissos, criando a variabilidade apropriada nos múltiplos valores imputados. Para efeitos deste trabalho, foi usada a biblioteca *mice* do R para proceder às IM selecionadas neste trabalho. Um sumário dos principais passos seguidos na aplicação desta técnica, usando o *mice*, é ilustrado na Figura 7.

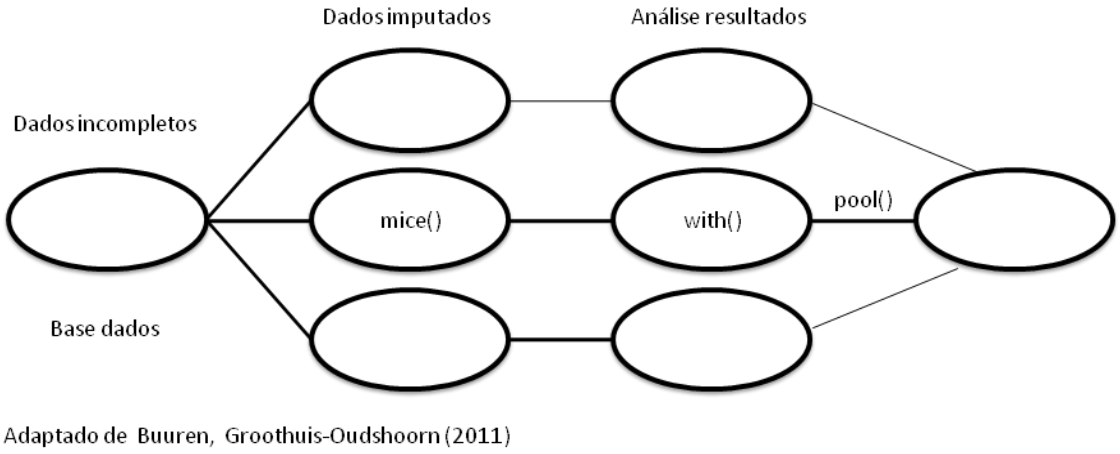


Figura 7 – Principais passos usados na Imputação Múltipla(45)

Existem três grandes vantagens da IM, relativamente à IS. Em primeiro lugar, quando as imputações são realizadas aleatoriamente, numa tentativa de representar a distribuição dos dados observados, a IM aumenta a eficiência da estimação. Em segundo lugar, as imputações obtidas a partir da IM representam recolhas de dados, feitas de modo aleatório e repetido, sob um modelo de não resposta, que permitem obter inferências válidas, combinando inferências de bases de dados completas, de forma direta. Esta característica permite que possam ser aplicados, de forma simples, métodos estatísticos habituais aos dados imputados. Em terceiro lugar, a geração de valores imputados aleatórios, sob mais do que um modelo, permite o estudo da sensibilidade das inferências a diferentes modelos de não resposta, usando métodos de dados completos repetidamente (2).

O procedimento base da IM é a substituição de cada valor omisso por M valores plausíveis. Cada valor é baseado na distribuição condicional da observação omissa, realizada de modo a que o conjunto de imputações represente, de forma apropriada, a informação dos valores omissos que está contida nos dados observados(3). São obtidas inferências válidas, já que se estão a calcular médias da distribuição dos dados omissos, dada a distribuição dos dados observados (46).

A IM consiste, basicamente, em **três passos** (9):

1. São obtidos M bancos de dados completos através de técnicas adequadas de imputação;

2. Separadamente, os M bancos são analisados por um método estatístico tradicional, como se realmente fossem conjuntos de dados completos;

3. Os M resultados encontrados no passo 2 são combinados para obter a chamada inferência da imputação repetida.

Supondo que há interesse em fazer inferências acerca do parâmetro k x 1 do vetor β do modelo substantivo (modelo sem dados omissos) e que temos possibilidade de fazer imputações Bayesianas apropriadas do modelo de imputação. Constroem-se *m* bases de dados completos. β^m e V^m são, respetivamente, a estimativa de β e a sua matriz de covariâncias gerada da m-ésima base de dados completa (m = 1, …, M). A estimativa da IM para β é a média das estimativas.

$$\beta^* = \frac{1}{M}\sum_{m=1}^{M}\beta^m$$

Também precisamos de uma medida da precisão de β* que reflita a incerteza das imputações. Uma grande vantagem prática da IM é a existência de uma expressão simples para a matriz de covariância de β* que pode ser aplicada de forma geral. Esta é conhecida como a fórmula de variância de Rubin e combina a variabilidade intra e entre-imputações de forma intuitiva. Define-se

$$W = \frac{1}{M}\sum_{m=1}^{M}V^m$$

como a média da matriz de covariâncias intra-imputações e

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\beta^m - \beta^*)(\beta^m - \beta^*)^t$$

como a matriz de covariância de β^m entre-imputações. Então, uma estimativa da matriz de covariância de β é dada por

$$V = W + \left(\frac{M+1}{M}\right) B.$$

Molenberghs apresenta uma justificação teórica da IM, com base na apresentação feita por Rubin (3). O cerne da IM é um argumento Bayesiano. Suponha que se tem um problema com dois parâmetros γ_1, γ_2 e dados y. Numa análise Bayesiana, estes têm uma distribuição conjunta *a posteriori*:

$$f(\gamma_1, \gamma_2 | \, y).$$

Suponha que o foco é em γ_2, sendo γ_1 considerado como *nuisance* (parâmetro de distúrbio). A distribuição *a posteriori* pode ser decomposto da seguinte forma

$$f(\gamma_1, \gamma_2 | \, y) = f(\gamma_1 | y) f(\gamma_2 | \gamma_1, y),$$

Pode ser demonstrado que a distribuição marginal *a posteriori* para γ_2 é expressa por:

$$f(\gamma_2 | \, y) = \, E_{\gamma_1}\{f(\gamma_2 | \gamma_1, y)\}.$$

Em particular, a média e variância *a posteriori* para γ_2 podem ser expressas por:

$$E(\gamma_2 | y) = \, E_{\gamma_1}\{E_{\gamma_2}(\gamma_2 | \gamma_1, y)\},$$

$$var(\gamma_2 | y) = \, E_{\gamma_1}\{var_{\gamma_2}(\gamma_2 | \gamma_1, y)\} + \, var_{\gamma_1}\{E_{\gamma_2}(\gamma_2 | \gamma_1, y)\}.$$

Estas podem ser aproximadas, usando momentos empíricos. Sejam $\gamma_1^m, m = 1, \ldots, M$, tiragens do modelo marginal *a posteriori* de γ_1. Então, aproximadamente,

$$E(\gamma_2 | y) \, \cong \, \frac{1}{M} \sum_{m=1}^{M} \{E_{\gamma_2}(\gamma_2 | \gamma_1^m, y)\} = \tilde{y}_2$$

e

$$var(\gamma_2 | y) = \frac{1}{M} \sum_{m=1}^{M} var_{\gamma_2}(\gamma_2 | \gamma_1^m, y) + \frac{1}{M-1} \sum_{m=1}^{M} \{E_{\gamma_2}(\gamma_2 | \gamma_1^m, y) - \, \tilde{y}_2\}^2$$

Esta formula pode ser generalizada para parâmetros vetoriais. A ligação final entre estas expressões e o procedimento de IM é usar γ_2 para representar parâmetros do modelo substantivo (sem valores omissos) e γ_1 para representar os valores omissos.

Eficiência

O principal atrativo da IM é a sua eficiência, mesmo quando o número de imputações M é reduzido. Em muitas aplicações, 3 a 5 imputações são suficientes para obter excelentes resultados. Rubin(2) mostra que a eficiência de uma estimativa baseada em M imputações é aproximadamente

$$\left[1 + \frac{\gamma}{M}\right]^{-1}$$

onde γ é a fração de informação omissa (FMI: fraction of missing information) para a quantidade a ser estimada. A fração γ quantifica o quão mais precisa a estimativa poderia ter sido, caso não existissem dados omissos. A Tabela 4 mostra as eficiências atingidas para vários valores de M e os respetivos valores de informação omissa. Esta tabela mostra que os ganhos rapidamente diminuem, após as primeiras imputações. Em muitas situação, é pouco vantajoso produzir e analisar mais do que poucas bases de dados imputadas (3). Há exceções a esta regra, que podem ser encontradas descritas em Carpenter(47).

Tabela 3– Eficiência relativa (em percentagem) da estimação por IM por número de imputações M e fração de informação omissa γ(3)

			γ		
M	0.1	0.3	0.5	0.7	0.9
2	95	87	80	74	69
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Escolha dos modelos de imputação múltipla

As especificações do modelo de imputação são a parte mais desafiante na IM. De acordo com Bureen (45), há sete grandes escolhas a serem feitas, das quais seis se aplicam ao presente trabalho.

1. Deve decidir-se se o pressuposto MAR é plausível. É um pressuposto seguro em muitos casos práticos, mas pode também ser suspeito. Na biblioteca *mice* lida com dados omissos do tipo MAR ou MNAR, mas a condição MNAR exige modelação adicional. Em suma, o investigador deve tentar compreender as razões da omissão dos dados (7).
2. A segunda escolha recai sobre a forma do modelo de imputação. A forma engloba a parte estrutural do modelo e a distribuição assumida dos resíduos. Na biblioteca *mice* esta forma tem de ser especificada para cada variável a ser imputada e varia de acordo com o tipo de variável (contínua, categórica, binária).
3. A terceira escolha prende-se com as variáveis a incluir como preditoras no modelo de imputação. O concelho genérico é incluir tantas variáveis relevantes quanto possível, incluindo as suas interações. No entanto, isto pode tornar-se moroso e complicado de gerir.
4. A quarta escolha é se se deve imputar variáveis que são funções de outras variáveis incompletas. Muitas bases de dados incluem variáveis transformadas, *scores* de somas, variáveis de interação, rácios, etc. Pode ser útil incorporar as variáveis transformadas no modelo de imputação.
5. A quinta escolha é relativa à ordem pela qual as variáveis devem ser imputadas. Há diferentes estratégias que podem ser usadas.
6. Diz respeito ao número de iterações que vão ser feitas na IM. A convergência deve ser monitorizada e isto pode ser feito de muitas formas, no *mice*.
7. A sétima e última escolha refere-se a *m*, o número de bases de dados constituídas na IM. Atribuir a *m* um valor baixo pode levar a um erro de simulação grande, especialmente se a fração de informação omissa é alta.

Complicações da IM

Escolha de variáveis no processo de imputação

Frequentemente, objetivo de uma análise estatística é estudar a associação entre um ou mais preditores (variáveis dependentes) e um *outcome*, mas alguns preditores apresentam valores omissos. Neste caso, o *outcome* possui informação acerca dos valores omissos dos preditores e esta informação deve ser usada (46). De facto, o

conjunto de variáveis usadas no aumento dos dados deve incluir todas as variáveis que serão usadas na análise planeada. Poderão incluir-se variáveis extra, que estejam altamente relacionadas com as variáveis a imputar. Isto trará mais precisão aos resultados da imputação e uma diminuição dos erros padrão (EP) (48). Outra questão poder-se-á levantar, quando é incluída a variável dependente da análise posterior para imputar valores omissos das variáveis independentes. Embora possa parecer, à primeira vista, que a utilização da variável dependente no modelo de imputação pode produzir coeficientes particularmente altos, é essencial incluí-la para garantir que os resultados das estimativas dos coeficientes de regressão não são enviesados. Com a imputação determinística, pode acontecer que os coeficientes sejam inflacionados, mas a introdução de um componente aleatório equilibra esta tendência e produz estimativas não enviesadas (48). É importante ter em conta que, por vezes, também existem valores omissos na variável dependente da análise, além das variáveis independentes. Alguns autores recomendam não proceder a imputação de dados omissos, quando estes existem na variável dependente (49). Estes casos deverão ser eliminados. No entanto, se existirem valores omissos na variável dependente e nas variáveis indepententes, então a primeira contém informação importante para contribuir para o cálculo dos coeficientes de regressão e, por esse motivo, os casos omissos não deverão ser imediatamente eliminados (48).

Variáveis sem distribuição normal

Muitos procedimentos de IM assumem que a variável a ser imputada apresenta uma distribuição normal, então incluir variáveis com distribuição não normal pode levar ao enviesamento dos resultados. Estas variáveis poderão ser transformadas para se aproximarem a uma variável normal antes da imputação e depois transformadas novamente, à sua escala inicial. Problemas diferentes se levantam relativamente a variáveis categóricas ou binárias e a melhor forma de tratar os dados omissos, neste caso é alvo de investigação (50). No entanto, é possível encontrar na biblioteca *mice* tipos de imputação que se adaptam a estes dados e são mais adequados para variáveis categóricas e binárias, como é o caso da regressão logística multinomial (45).

Plausibilidade do pressuposto MAR

O pressuposto MAR não é uma propriedade dos dados, mas sim uma justificação para a escolha de determinadas análises. Por outro lado, não se pode provar que os dados são MAR. Apenas se pode especular acerca do assunto (12). Assim, é sensato incluir uma grande variedade de variáveis nos modelos de imputação, incluído todas as variáveis na análise substantiva e, se for possível, todas as variáveis preditoras de valores omissos disponíveis e todas as variáveis que influenciam o processo gerador dos dados omissos, se for conhecido, mesmo que não sejam do interesse da análise substantiva. Não o fazer, pode implicar que o pressuposto MAR não seja plausível.

Dados omissos que não são MAR/MCAR

Alguns dados omissos são, inerentemente, MCAR já que não é possível ter em conta diferenças sistemáticas entre casos com valores omissos e valores observados. Nesta situação a utilização da IM pode levar a resultados erróneos e possivelmente mais enviesados do que uma análise de casos completos.

Interações e não linearidade na IM

Os métodos de imputação múltipla são satisfatórios para estimar os efeitos das variáveis com valores omissos, mas apresentam maior dificuldade quando se pretende estimar efeitos de interações. O problema surge visto que, embora o modelo normal multivariado seja adequado para imputar valores que reproduzem associações lineares entre variáveis, não modela momentos de ordens superiores. Podem existir duas soluções para esta questão: a variável de interação ou a variável transformada é criada antes da imputação ou, caso pretendamos uma interação, em que uma das variáveis é binária, criam-se duas bases de dados (uma para cada categoria da variável binária) e procede-se a imputações separadas. Quando o processo está concluído, as bases de dados podem ser recombinadas numa só, criando então a variável de interação, posteriormente. Ambos os métodos parecem ter resultados satisfatórios para contornar esta questão (48).

4.4.1. Imputação múltipla por *predictive mean matching*

A imputação por PMM, como IM, baseia-se nos mesmos princípios que a anteriormente descrita, exceto que em vez de se gerar apenas um valor, são gerados vários. A variância estatística entre e intra bases de dados imputadas é então usada

para incorporar a variância de imputação nos calculos e obter estimativas mais precisas dos erros padrão e testes de significância.

4.4.2. Imputação múltipla por regressão linear não Bayesiana

A regressão linear não Bayesiana é um método estocástico que consiste na utilização de variáveis auxiliares como preditoras da variável de interesse, a ser imputada. Considerando o modelo de regressão linear, a imputação é feita através da equação,

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{i,j}, i = 1,2, \dots, n$$

para todos os valores $y_i$omissos.

As variáveis auxiliares podem ser qualitativas ou quantitativas. Este método apresenta alguns incovenientes – em primeiro lugar, é necessário especificar o modelo de regressão mais adequado. Depois, como os valores imputados são coincidentes com o modelo de regressão, as covariâncias e as correlações entre as variáveis apresentam-se inflacionadas. Além disso, como não usa valores observados da amostra, pode imputar valores pouco realistas (com, por exemplo, a idade negativa). No entanto, produz um menor enviesamento do que outros métodos determinísticos, como é o caso da imputação pela mediana. Esta função, na biblioteca *mice*, não incorpora a variabilidade do peso da regressão e, por isso, não é uma imputação apropriada, de acordo com as regras de Rubin. Para amostras pequenas, a variabilidade dos dados imputados é, por isso mesmo, subestimada (51). O presente estudo apresenta uma amostra de dimensão relativamente grande, o que poderá implicar que o enviesamento deste tipo de IM seja menor. De qualquer forma, a escolha desta análise como técnica de IM prendeu-se com a comparação entre a mesma e o PMM, de modo a estabeler comparações.

4.5. Seleção das variáveis associadas ao IMC, nos modelos múltiplos

Realizou-se uma análise bivariada entre IMC e as variáveis explicativas, no modelo de casos completos original. Recorreu-se ao teste de correlação de Pearson, para variáveis contínuas, ao teste T para variáveis binárias, e à ANOVA seguida do teste Tukey, para variáveis com mais de 2 categorias. Independentemente do resultado, todas as variáveis presentes na base de dados, foram incluídas no modelo completo,

visto existir interesse epidemiológico em compreender a associação das variáveis explicativas ao IMC, no modelo múltiplo. Foi usado um modelo linear generalizado, com função de ligação identidade. A partir daí, foi usada um seleção *backwards* para definir as variáveis presentes no modelo final. O teste utilizado para a retirada da variável, em cada passo do procedimento backwards, foi o teste da razão de verosimilhanças. O valor-p máximo usado para a variável independente entrar no modelo final é 0.10. Este procedimento foi repetido para todos os modelos calculados.

No caso da IM, utilizou-se um procedimento semelhante. Foi realizada uma seleção de variáveis *backwards*, através do teste de Wald, para comparação entre modelos. Cada par de modelos comparados, diferia apenas numa variável, para que a seleção pudesse ser feita. Esta variável era escolhida de acordo com o maior valor-p do modelo *pooled*, produzido após a IM. Mais uma vez, a seleção era interrompida, quando a variável apresentava um valor-p inferior a 0.10.

A variável a ser imputada (escolaridade) tem, obrigatoriamente, de estar no modelo final, visto que é essa a única diferença entre os modelos de regressão, obtidos após aplicação de cada técnica de imputação. O modelo de casos completos original foi também avaliado, relativamente à qualidade de ajustamento dos dados. Para esse fim, utilizaram-se métodos gráficos, recorrendo ao calculo do *leverage*, dos resíduos padronizados e da distância de Cook.

Todos os modelos foram testados para garantir a não existência de multicolinearidade entre variáveis selecionadas, através do calculo do *variance inflaction factor* (VIF). Os resultados são apresentados em detalhe apenas no primeiro modelo, de casos completos original.

4.6. Modelo de imputação

Seguindo os príncipios de Buuren (45), os vários modelos de IM seguiram as seguintes escolhas:

1. A omissão em escolaridade foi considerada do tipo MAR, por parecer plausível que assim fosse, de acordo com os resultados da exploração de dados omissos
2. A opção da forma do modelo de imputação prendeu-se com o facto de escolaridade e idade serem variáveis contínuas, sendo por isso o PMM e a RLN,

métodos apropriados. Enquanto que o PMM foi escolhido também enquanto técnica de IS, tendo a vantagem de poder ser comparado na IM, a RLN serviu como exemplo de um tipo de modelo menos interessante e que tipo de resultados podem surgir a partir daí

3. É aconselhado na literatura que se escolham o maior número de preditores possíveis para o modelo de imputação, de modo a tornar o pressuposto de MAR mais plausível. Este tipo de estratégias podem ser denominadas de inclusivas, visto que há um uso liberal das variáveis selecionadas (52). No presente estudo, optou-se por esta esta estratégia, incluído todas as variáveis presentes na base de dados no modelo de imputação. Deve ser salientado, no entanto, que a base de dados original continha variáveis originais, como *snack*, e variáveis recategorizadas da variável original, como *snackr*. Neste caso, visto não fazer sentido incluir ambas no modelo, optou-se pelas variáveis recategorizadas, tendo-se feito uma pré-seleção dos preditores da variável a ser imputada.
4. A avaliação da convergência das iterações foi feita por meios gráficos, disponíveis na biblioteca *mice*, e encontra-se em anexo neste trabalho.
5. Foi escolhido m = 5, como o número de imputações a serem realizadas, em todos os casos. Este é um valor habitualmente usado na literatura e com a possibilidade de obtenção de excelentes resultados (45,53).

4.7. Medidas de comparação

Para efeitos de comparação dos modelos, recorreu-se a sete parâmetros:

1. Seleção de variáveis, a partir do modelo completo, para o modelo final
2. Direção da estimativa dos coeficientes da regressão
3. Valor absoluto da estimativa dos coeficientes da regressão
4. Erro padrão da estimativa dos coeficientes da regressão
5. Valor-p associado a cada estimativa dos coeficientes de regressão
6. AIC dos modelos
7. Coeficiente de determinação, R^2, dos modelos obtidos

O R^2 dos modelos obtidos, visto tratarem-se de modelos lineares generalizados, foi calculado a partir da seguinte fórmula (54):

$$\rho^2 = 1 - \frac{GL\ Função\ desvio\ reduzida}{GL\ Função\ desvio\ nula} \times \frac{Função\ desvio\ reduzida}{Função\ desvio\ nula}$$

No caso concreto dos modelos obtidos após IM, esta medida foi calculada a partir do R^2 de cada um dos modelos obtidos em cada uma das bases de dados imputadas, após o processo de seleção de variáveis, no modelo conjunto. O R^2 final é o resultado da média dos cinco R^2 otidos. O mesmo processo foi usado para o calculo do AIC, do modelo linear generalizado após a IM.

5. Resultados

5.1. Análise dos casos completos e determinantes do Índice de Massa Corporal nos imigrantes

5.1.1. O índice de massa corporal

A Figura 8 apresentao histograma da variável IMC, na amostra selecionada e permite-nos verificar, tal como é comum, que há mais probabilidade de encontrar pessoas com excesso de peso na amostra (cauda direita) em comparação com pessoas com baixo peso (cauda esquerda).

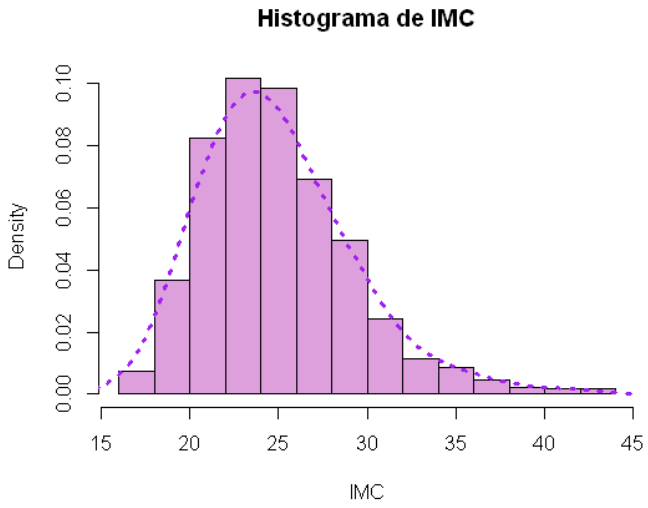


Figura 8 - Distribuição do IMC dos imigrantes do projeto SAIMI (n = 1980)

É sabido que a distribuição do IMC se altera de maneira característica. À medida que a média aumenta, o seu desvio padrão também aumenta. O que isto indica é que com o aumento das populações na média de IMC, a sua distribuição também se alarga de modo específico. Quando a população encontra um ambiente que favorece a obesidade, toda a população é afetada. Ao mesmo tempo, a suscetibilidade individual a este ambiente difere e aqueles que estão na cauda direita crescem muito mais no IMC, do que aqueles que estão na cauda esquerda (55). Ou seja, os valores de IMC superiores à média são encontrados em maior número do que aqueles inferiores à mesma. Assim e, de forma a otimizar a análise estatística dos dados, optou-se por usar um modelo linear generalizado cuja variável resposta tem distribuição gama. A Figura 8

é ilustrativa da explicação anterior e comprova a distribuição do IMC, com uma cauda direita mais pesada.

4.1.3. Análise bivariada do índice de massa corporal nos imigrantes

Para cada uma das variáveis presentes no estudo, previamente selecionadas das variáveis recolhidas no âmbito do projeto SAIMI, devido à sua relevância clínica e possível influência na explicação do IMC, apoiada pela revisão da literatura, foram aplicados os testes bivariados adequados.

A idade está positivamente associada ao IMC, de forma significativa (coeficiente linear de Pearson, r = 0.338; valor-p < 0.001) (Figura 9).

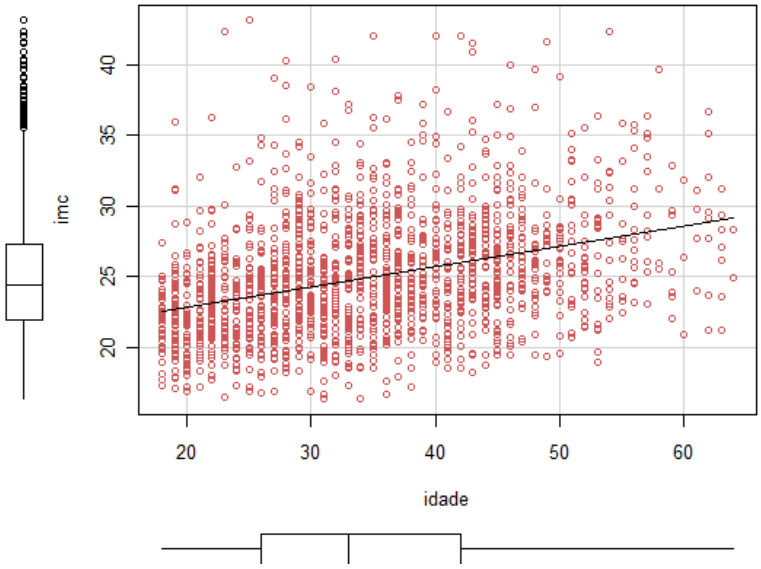


Figura 9 – Associação entre o IMC e a idade

A associação entre IMC e anos de residência em Portugal é também uma associação positiva e significativa (r = 0.259; valor-p < 0.001), ilustrada na Figura 10.

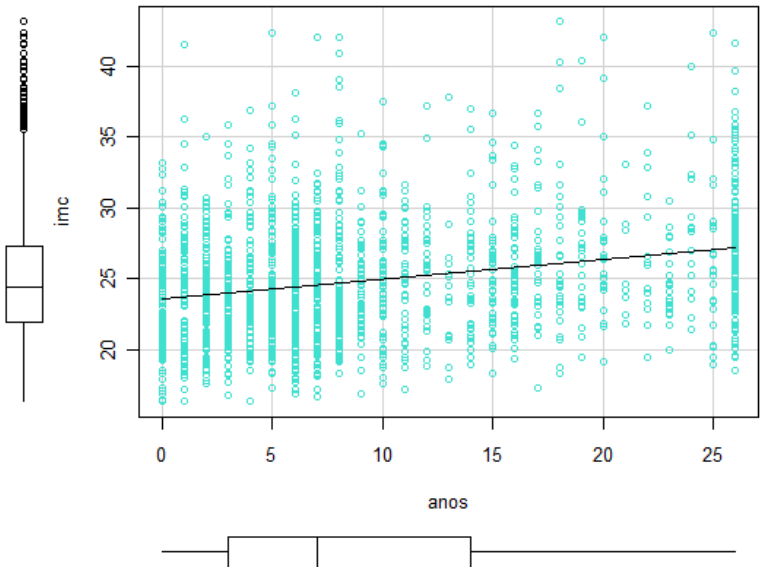


Figura 10– Associação entre IMC e anos de residência em Portugal

O IMC aparenta ter uma associação negativa, mas não significativa com os anos de escolaridade completos pelos imigrantes (r = -0.108; valor-p <0.001) (Figura 11).

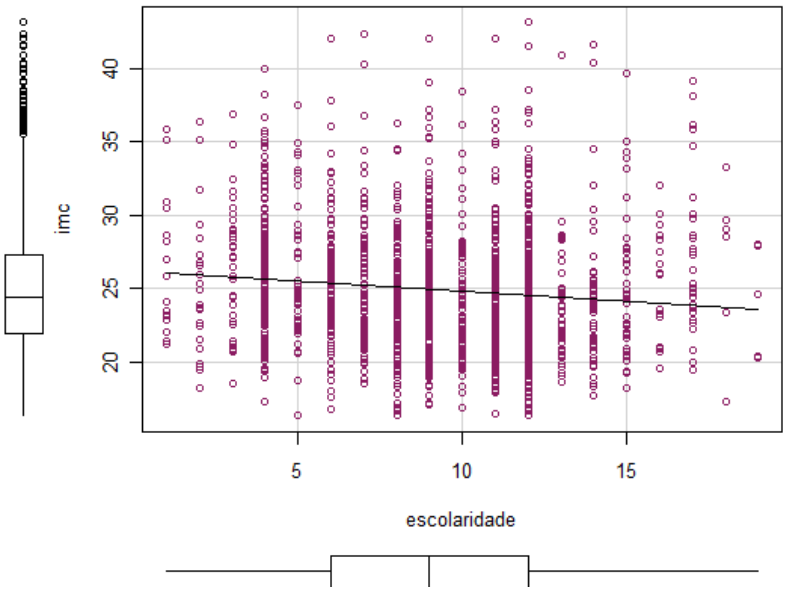


Figura 11 – Associação entre IMC e anos de escolaridade completos

Relativamente às associações de IMC com as variáveis categóricas, verificou-se uma associação entre o mesmo e o estado civil, através da ANOVA (valor-p < 0.001). Aplicou-se o teste Tukey e encontraram-se diferenças significativas entre as médias de IMC dos solteiros e casados e dos solteiros comparativamente a outro estado civil (divorciados ou viúvos), ao nível de significância de 1%. Neste caso, a média é superior nos casados e outros. Os casados e os classificados na categoria de outro estado civil

não apresentam diferenças significativas na média de IMC (valor-p = 0.66). A Figura 12 ilustra as diferenças no IMC através das medianas.

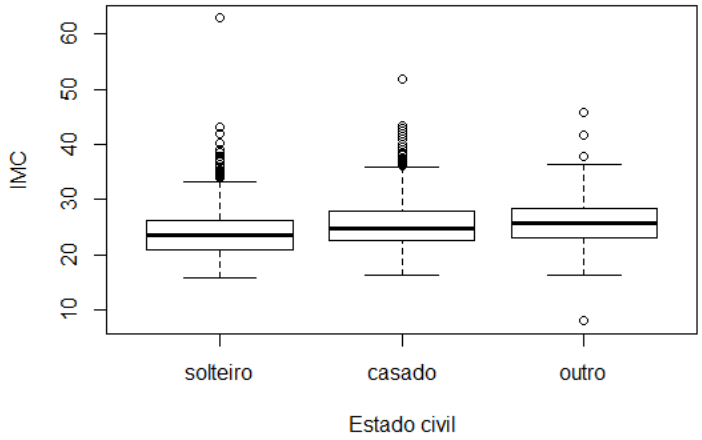


Figura 12 - *Boxplot* do IMC por estado civil

O IMC dos imigrantes difere, de forma significativa, pela sua origem – brasileira ou africana (valor-p < 0.001). Encontrou-se um IMC médio menor nos imigrantes brasileiros (24.3 kg/m^2), comparativamente aos imigrantes africanos (25.5 kg/m^2).

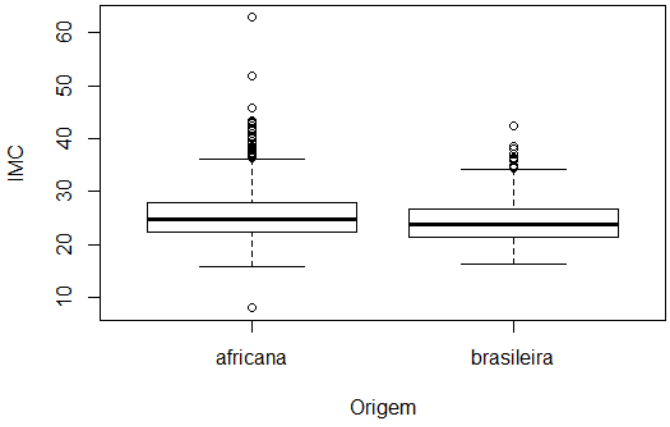


Figura 13 - *Boxplot* do IMC por origem dos imigrantes

A variável número de refeições principais ingeridas por dia não apresentou diferenças significativas entre as médias de IMC (valor-p = 0.201) (Figura 14).

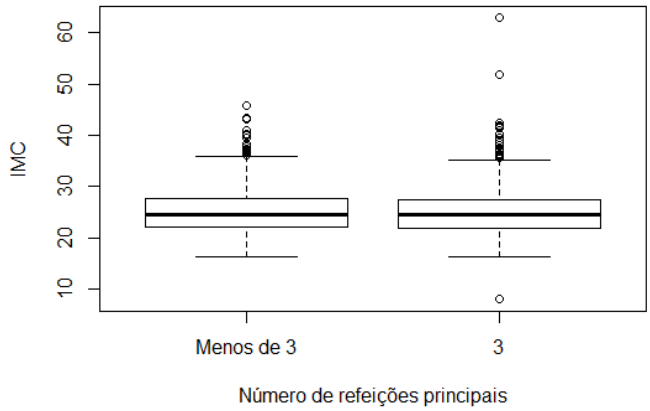


Figura 14 – *Boxplot* do IMC por número de refeições principais

A variável número de refeições intermédias ingeridas por dia apresentou diferenças significativas entre as médias de IMC dos diferentes grupos (valor-p < 0.001). De acordo com o teste Tukey, as categorias de refeições intermédias que apresentam diferenças estatisticamente significativas entre as respetivas médias de IMC são Dois-Zero (p = 0.020); Mais de três-Zero (p = 0.004); Dois-Um (p = 0.034); Mais de três-Um (p = 0.007) (Figura 15).

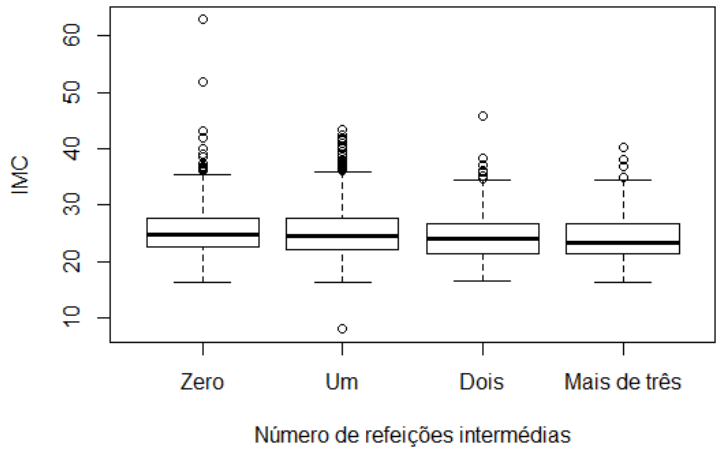


Figura 15 – *Boxplot* do IMC por número de refeições intermédias

4.1.4. Análise múltipla dos fatores determinantes do índice de massa corporal

Adotou-se o modelo linear generalizado gama com variável resposta IMC.

O método de seleção de variáveis para inclusão no modelo foi o procedimento sequencial *backwards*. Apresenta-se abaixo a tabela ilustrativa dos resultados obtidos,

em cada passo do procedimento (Tabela 5). O resultado da estimação do modelo final encontra-se na Tabela 7.

O modelo 1 inicial foi:

$$E(IMC) = \beta_0 + \beta_1 * Sexo_{masculino} + \beta_2 * Idade + \beta_3 * Anos + \beta_4 * Escolaridade + \beta_5 * Estado\ civil_{casado} + \beta_6 * Estado\ civil_{outro} + \beta_7 * Refeições_{3\ refeições} + \beta_8 * Snack_{Um} + \beta_9 * Snack_{Dois} + \beta_{10} * Snack_{Três\ ou\ mais} + \beta_{11} * Origem_{brasileira}$$

Tabela 4 – Resultados do modelo de seleção *backwards* no modelo de regressão com casos completos (Cenário 1)

Variável	Modelo	Deviance	valor-p
-	1	43.127	
Sexo	2	43.141	0.461
Origem	4	47.962	0.341
Refeições	5	48.078	0.124

As variáveis sexo, origem e refeições foram eliminadas do modelo final que é apresentado na tabela que se segue. A seguir, procedeu-se a avaliação da multicolinearidade das variáveis selecionadas para o modelo final, através do cálculo do VIF (*Variance Inflation Factor*). Todos os valores de VIF eram inferiores a 2 (Tabela 6), tendo-se mantido as variáveis selecionadas no modelo.

Tabela 5 – Resultados do calculo do VIF para as variáveis modelo de regressão finalcomcasos completos (Cenário 1)

Variáveis	VIF
Idade	1.615
Estado civil (Casado)	1.277
Estado civil (Outro)	1.209
Anos	1.343
***Snack* (Um)**	1.478
***Snack* (Dois)**	1.411
***Snack* (Três ou mais)**	1.291
Escolaridade	1.132

As variáveis mantidas no modelo final foram a idade, o estado civil, os anos de residência no país e o número de *snacks* consumidos ao longo do dia, além da escolaridade. Por cada ano adicional na idade e nos anos de residência no país, aumenta-se em média o IMC em 0.104 kg/m^2 e 0.078 kg/m^2. Os sujeitos casados ou de outro estado civil têm em média mais 0.601 e 0.337 kg/m^2 em IMC, do que os sujeitos

solteiros. Quem consome um, dois ou três ou mais *snacks* ao longo do dia tem em média menos 0.142, 0.696 e 0.902 kg/m^2 de IMC, comparativamente a quem não consome *snacks* (Tabela 7).

Tabela 6 – Estimativas do modelo linear generalizado gama do IMC com casos completos (Cenário 1)

Covariável	Estimativa	Erro padrão	Valor t	p-value
Ordenada na origem	20.411	0.381	54.328	< 0.001
Escolaridade	0.022	0.030	0.756	0.450
Idade	0.104	0.012	8.613	< 0.001
Estado civil (Casado)[+]	0.601	0.219	2.743	0.006
Estado civil (Outro)[+]	0.337	0.451	0.747	0.455
Anos	**0.078**	0.014	5.381	< 0.001
***Snack* (Um)**[a]	-0.142	0.244	-0.582	0.561
***Snack* (Dois)**[a]	-0.696	0.276	-2.522	0.012
***Snack* (Três ou mais)**[a]	-0.902	0.329	-2.740	0.006
AIC	9856.6			
R^2	15.2%			

[+] Categoria referência: solteiro
[a] Categoria referência: Zero *snacks*

A equação do modelo final estimado é a seguinte:

$$\mathrm{E(IMC)} = 20.411 + 0.022 \times \mathrm{Escolaridade} + 0.104 \times \mathrm{Idade} + 0.601 \times \mathrm{Estado\ civil}_{\mathrm{Casado}} + 0.337 \times \mathrm{Estado\ civil}_{\mathrm{Outro}} + 0.078 \times \mathrm{Anos} - 0.142 \times \mathrm{Snack}_{\mathrm{Um}} - 0.696 \times \mathrm{Snack}_{\mathrm{Dois}} - 0.902 \times \mathrm{Snack}_{\mathrm{Três\ ou\ mais}} \quad [2]$$

Para avaliar a qualidade do ajuste do modelo, procedeu-se a uma análise gráfica dos resíduos do mesmo, assim como dos possíveis pontos influentes. Procedeu-se também à verificação da relação linear entre as variáveis explicativas contínuas e os resíduos do modelo (Anexo 9.3).

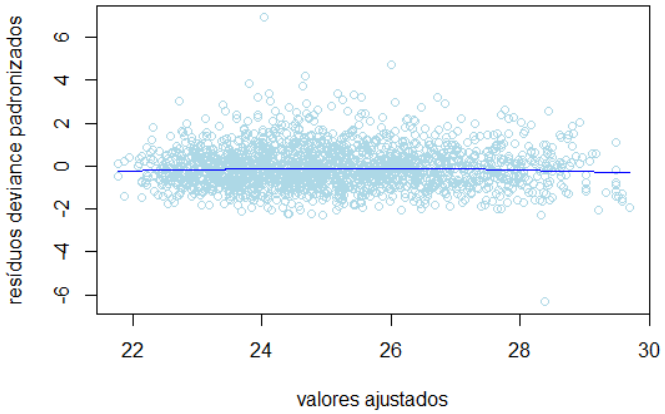


Figura 16 – Resíduos de deviance padronizados contra valores ajustados do modelo final [2]

Os resíduos parecem ter uma variância constante, uma média próxima de zero e não apresentam nenhum padrão discernível, atendendo assim ao pressuposto de homocedasticidade (Figura 16).

Como podemos ver na Figura 17 existem vários pontos potencialmente influentes (com a razão entre *leverage* (h_{ii}) e p/n, onde p é o número de parâmetros do modelo e n o número de observações (56), superior a 2), mas apenas aqueles que tenham um resíduo correspondente elevado é que são preocupantes. Para se aferir se existiam pontos nesta situação, foi calculada a distância de Cook (Figura 18) e produziu-se o gráfico baseado no valor de *leverage versus* os resíduos de deviance padronizados (Figura 19).

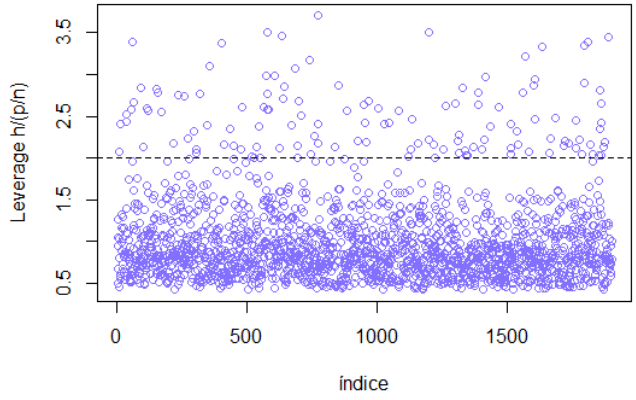


Figura 17– Possíveis pontos influentes do modelo final [2]

O *leverage* de uma observação y_i é dado por h_{ii} e mede a influência de y_i em $\hat{\mu}$. Quanto maior for h_{ii}, maior é o peso que uma observação y_i tem no valor ajustado. A

Figura 18 mostra que a distância de Cook é muito baixa para a maioria dos pontos, no entanto podemos ver que existem pontos cuja razão entre *leverage* (h_{ii}) *e p/n,* é superior a 2 e podem, por isso, ser possíveis pontos influentes (Figura 17).

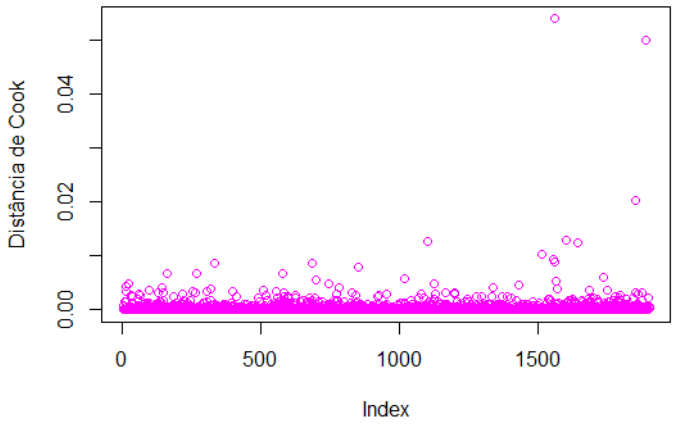


Figura 18– Distância de Cook

A Figura 19 apresenta os resíduos *deviance* padronizados *versus* o cálculo do *leverage*. Podemos desconfiar da existência de pontos influentes, caso estes se encontrem nos quadrantes superior e inferior do lado direito da Figura 19 (ou seja, observações cujo *leverage* e o resíduo padronizado, em módulo, sejam elevados). Encontramos um ponto nesta situação. Experimentou-se retirar esta observação da análise, o que não produziu qualquer efeito nos resultados obtidos, tendo-se optado por reinserir a mesma.

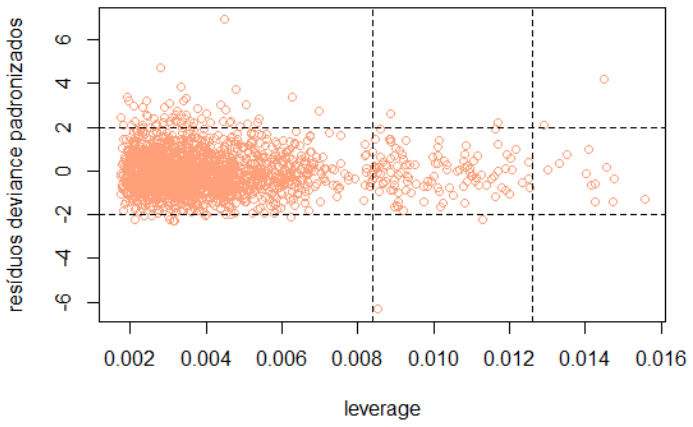


Figura 19 - Pontos influentes do modelo final [2]

5.2. Caracterização de dados omissos na variável escolaridade

Podemos verificar através da figura abaixo que a variável escolaridade é aquela que apresenta maior proporção de valores omissos, de entre todas as variáveis explicativas incluídas no modelo final da equação [2]. Escolaridade tem cerca de 7% de dados omissos, seguida de número de refeições intermédias (*snack*) com aproximadamente 3% e anos de residência em Portugal (anos) e número de refeições principais com valores à volta de 1%. A variável estado civil apresenta dados omissos perto de 0% e as restantes variáveis não apresentam dados omissos (Figura 20).

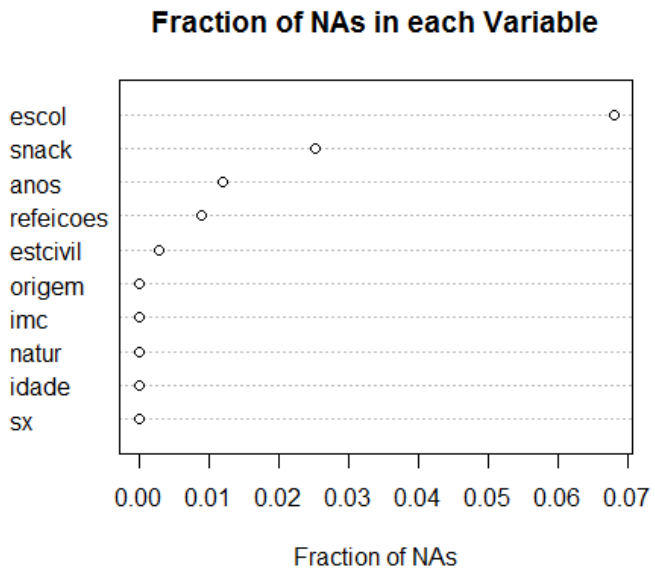


Figura 20- Fração de dados omissos por variável

O gráfico que se segue apresenta uma análise de *clusters* hierárquica que permite revelar combinações de dados omissos, entre variáveis (10). Assim, podemos verificar que as variáveis escolaridadee anos de residência em Portugal tendem a estar omissas nos mesmos sujeitos, assim como as variáveis número de refeições principais e intermédias. O eixo das ordenadas observado na Figura 21 corresponde à fração de observações para as quais ambas as variáveis estão omissas.

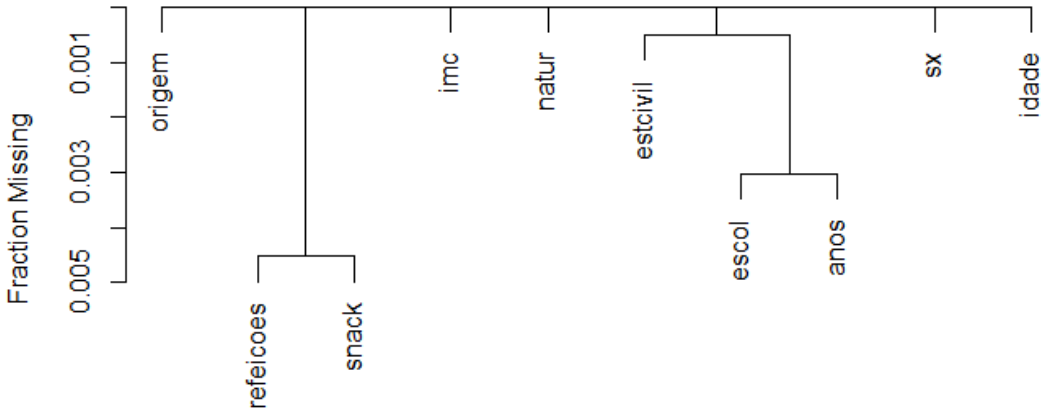


Figura 21 – Análise de *clusters* hierárquica dos dados do projeto SAIMI (n = 1980)

A Figura 22 mostra o resultado da árvore de regressão, uma técnica denominada por Harrell (10) de *recursive partitioning*, para compreender o padrão dos sujeitos com dados omissos em escolaridade. Pode verificar-se que a primeira partição dá-se pela separação dos sujeitos com IMC inferior a 29.18 kg/m^2. A segunda partição é feita dentro dos indíviduos com IMC inferior a 29.18, separando-se dos que têm IMC superior ou igual a 23.13 kg/m^2. As partições são feitas sucessivamente até a condição de paragem ser atingida, o que neste caso significa ter idade inferior a 4.5 anos. Ou seja, o padrão mais forte, encontrado por esta função, prende-se com a variável idade.

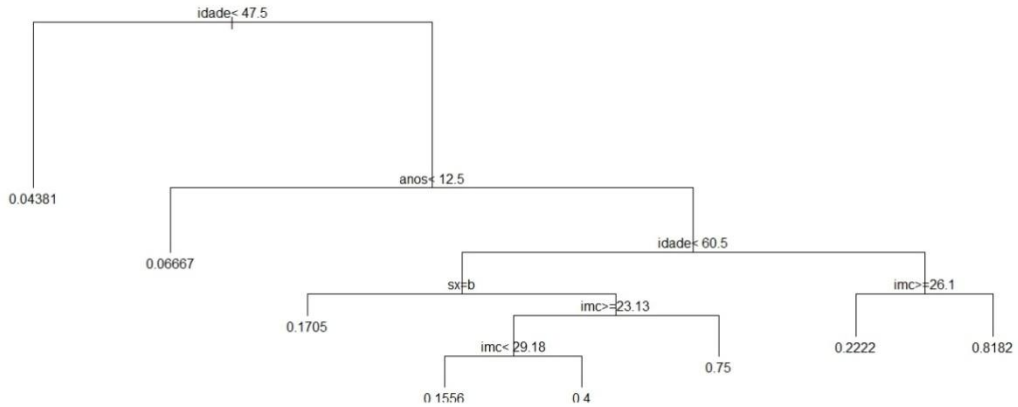


Figura 22– Árvore de regressão dos dados do projeto SAIMI (n = 1980)

Na Figura que se segue, caracteriza-se a amostra de acordo com o facto de apresentar ou não valores omissos para a variável escolaridade. Pode constatar-se que as mulheres têm maior percentagem de dados omissos de escolaridade do que os homens. Os sujeitos com outro estado civil que não solteiro ou casado, africanos e aqueles que fazem zero ou um *snacks* por dia apresentam também maior percentagem de dados omissos nesta variável. No geral, os sujeitos com maior percentagem de dados omissos noutras variáveis, que não a escolaridade, têm maior percentagem de dados omissos na escolaridade (Figura 23).

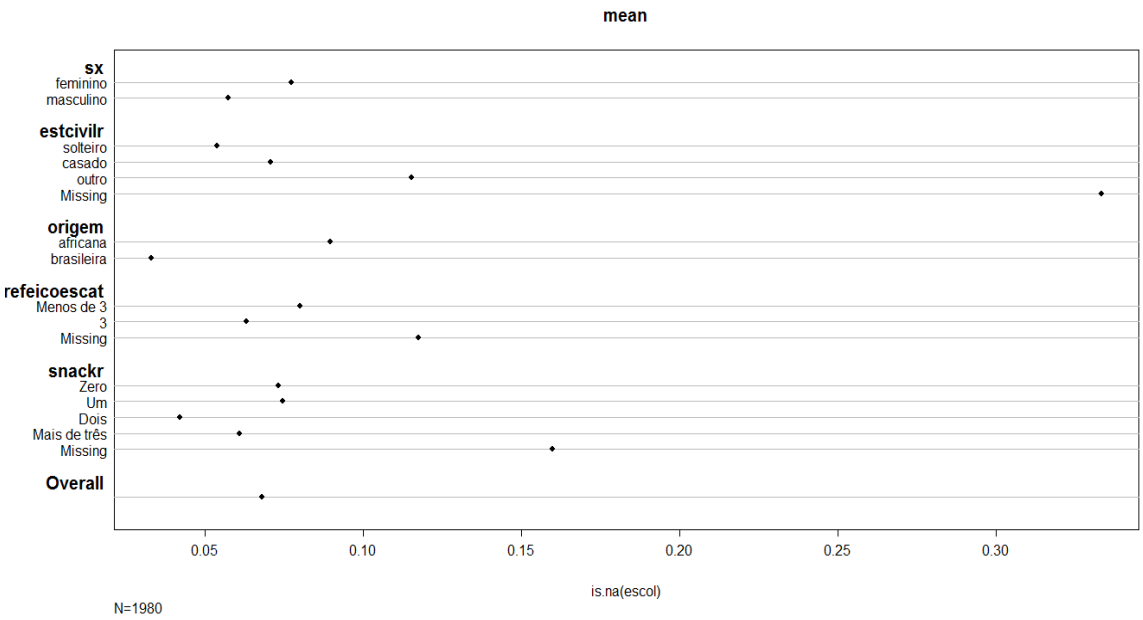


Figura 23 – Descrição univariada da proporção de sujeitos com dados omissos na variável escolaridade dos dados do projeto SAIMI (n = 1980)

Obviamente, estas técnicas mostram-nos uma perspetiva uni/bivariada. Existem mais técnicas gráficas para este fim (Anexo 9.4). No caso do modelo múltiplo, de regressão logística, cuja variável resposta é ter ou não dados omissos na variável escolaridade (*is.na(escol)*) pode ajudar a perceber se as associações sugeridas anteriormente se mantêm na análise múltipla (Tabela 8).

Tabela 7– Estimativas do modelo de regressão logística com variável resposta is.na(escol)[+]

Covariável	Estimativa	Erro padrão	OR	Valor Z	p-value
Ordenada na origem	-4.647	0.691	0.010	-6.723	< 0.001
Sexo (Masculino)	-0.532	0.211	0.588	-2.518	0.0118
Idade	0.067	0.011	1.069	6.067	< 0.001
Origem (Brasileira)	-0.562	0.319	0.570	-1.759	0.079
Anos	0.032	0.015	1.033	2.173	0.030
AIC	792.46				
R^2	13.3%				

[+] Modelo ajustado para idade, estado civil, IMC, número de refeições principais e número de refeições intermédias. As variáveis apresentadas são as que apresentam uma associação significativa com a variável resposta

Podemos verificar que as variáveis sexo, idade, origem e anos de residência em Portugal estão associadas, de forma estatisticamente significativa, com o nível de significância de 10%, com a variável resposta (is.na(escol)). Assim, podemos afirmar que ser homem, mais jovem, de origem brasileira e a residir há menos anos em Portugal está associado a um menor risco de ter escolaridade como valor omisso.

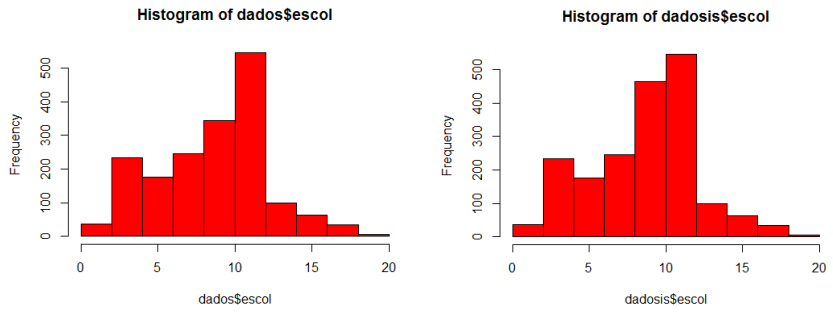
Isto poderá indicar que as variáveis não apresentam um padrão MCAR, visto que a omissão de escolaridade depende de variáveis observadas na investigação. Usualmente, a probabilidade de um dado ser omisso depende de outras variáveis observadas em relação ao sujeito, isto é, a razão para perda de informação pode ser baseada noutras informações observadas (8).

5.3. Cenário 1

5.3.1. Imputação simples pela substuição da mediana (Cenário 1)

A imputação simples (IS) realizada neste trabalho foi feita por substituição dos valores omissos na variável escolaridade pela mediana da escolaridade. A mediana foi escolhida, devido à distribuição da variável escolaridade, apresentada na Figura 24 (A), através do seu histograma. Pode constatar-se que se trata de uma distribuição assimétrica.

O efeito da imputação simples pela mediana na distribuição da variável escolaridade pode ser verificado através do gráfico (B), na Figura 24.



(A) (B)

Figura 24- Distribuição da variável escolaridade na base de dados completa e após imputação pela mediana. (A): dados orginais, n = 1777 . (B): dados imputados pela mediana, n = 1980.

A leitura da Figura 24 permite-nos perceber uma alteração na distribuição da variável escolaridade, após a imputação pela mediana. Os valores encontram-se mais centrados em torno da mediana e a variância diminuiu.

Apresenta-se abaixo a tabela dos resultados obtidos em cada passo da seleção de variáveis (Tabela 9).

Tabela 8 – Resultados da seleção *backwards* para os dados imputados pela mediana no modelo de linear generalizado gama do IMC (Cenário 1) (n = 1980)

Variável	Modelo	Deviance	valor-p
-	1	47.890	
Origem	2	47.927	0.340
Sexo	3	47.962	0.260
Refeições	4	48.027	0.124

As variáveis mantidas no modelo foram, idade, estado civil, anos e número de *snacks* consumidos por dia, além da escolaridade. Mais uma vez, foi utilizado o cálculo do VIF como diagnóstico de multicolinearidade. Todos os valores foram inferiores a 2, tendo-se mantido as variáveis selecionadas no modelo.

Tabela 9 – Estimativas do modelo linear generalizado gama do IMC para dados imputados pela mediana (Cenário 1) (n = 1980)

Covariável	Estimativa	Erro padrão	Teste T	P-value
Ordenada na origem	20.564	0.522	39.377	<0.001
Escolaridade	0.018	0.030	0.591	0.544
Idade	0.103	0.011	8.873	<0.001
Estado civil (Casado)[+]	0.478	0.215	2.217	0.026
Estado civil (Outro)[+]	0.237	0.436	0.543	0.587
Anos	**0.075**	0.014	5.287	<0.001
***Snack* (Um)**[a]	-0.170	0.239	-0.712	0.477
***Snack* (Dois)**[a]	-0.583	0.239	-2.132	0.033
***Snack* (Três ou mais)**[a]	-0.855	0.325	-2.632	0.009
AIC	10603			
R^2	14.4%			

[+] Categoria referência: solteiro
[a] Categoria referência: Zero *snacks*

As variáveis associadas ao IMC mantêm-se, relativamente ao modelo de regressão com dados completos. Os sujeitos mais velhos, casados ou viúvos/divorciados, que vivem há mais anos no país e que não consomem *snacks* ao longo do dia têm em média um IMC superior aos mais jovens, solteiros, que vivem há menos anos no país e que consomem um ou mais *snacks* ao longo do dia (Tabela 10).

5.3.2. Imputação simples por *predictive mean matching* (Cenário 1)

Um dos fatores primordiais na implementação do PMM é a capacidade de predição do mesmo, relativamente à escolaridade nos sujeitos omissos. Foi usada uma regressão linear para este efeito, cuja variável resposta era a escolaridade. Os resultados encontrados nesta regressão encontram-se descritos na (Tabela 11).

Tabela 10 – Estimativas do modelo de regressão linear com variável resposta escolaridade

Variáveis	Estimativa	Erro padrão	Teste T	Valor-p
Ordenada na origem	11.295	0.540	20.930	<0.001
Idade	-0.111	0.009	-11.823	<0.001
Sexo(masculino)	-0.099	0.157	-0.628	0.530
Estado civil(casado)[+]	-0.025	0.182	-0.137	0.891
Estado civil(outro)[+]	0.324	0.355	0.912	0.362
Origem(brasileira)	1.270	0.200	6.334	<0.001
***Snack*(Um)**[a]	0.190	0.194	0.976	0.329
***Snack*(Dois)**[a]	0.302	0.270	1.117	0.264
***Snack*(Três ou mais)**[a]	0.302	0.270	1.117	0.264
Refeições(Três)	0.449	0.177	2.541	0.011
Anos	0.011	0.013	0.801	0.423
IMC	0.024	0.019	1.271	0.204
R^2	14.37%			

[+] Categoria referência: solteiro
[a] Categoria referência: Zero *snacks*

A capacidade de predição do modelo é, infelizmente, fraca. Apresenta um R^2 de cerca de 14%. A Figura 25 ilustra esta constatação, com os dados ajustados da escolaridade no eixo das ordendase a escolaridade observada no eixo das abcissas. A leitura da figura pode ser feita da seguinte forma: um sujeito com dez anos de escolaridade observados, terá uma valor de anos de escolaridade ajustado pelo modelo entre os 6 e os 12. Esta limitação pode prender-se com o número reduzido de variáveis no modelo de regressão, que não são suficientes para predizer a escolaridade nos imigrantes. Contudo, é de ressaltar que o objetivo do projeto SAIMI não era a predição desta variável.

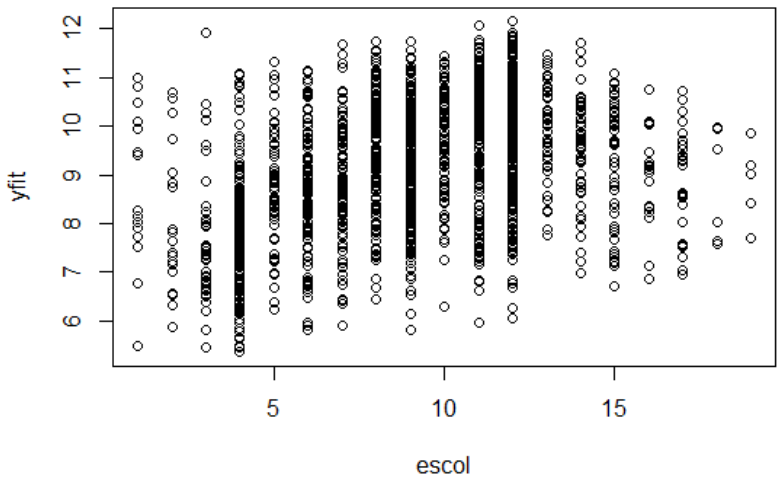


Figura 25 – Valores ajustados da escolaridade pelo modelo de regressão linear

O resultado do porcesso de seleção de variáveis, gerado a partir dos dados após IS por PMM é apresentado de seguida, na Tabela 12.

Tabela 11 – Resultados do modelo de seleção *backwards* no modelo linear generalizado gama do IMC com dados imputados pelo PMM (Cenário 1) (n = 1980)

Variável	Modelo	Deviance	valor-p
-	1	47.859	
Origem	2	47.887	0.315
Sexo	3	47.922	0.226
Refeições	4	47.962	0.124

Pode constatar-se, na Tabela 12 e 13, que as variáveis selecionadas são as mesmas que nos dois casos anteriores e as direções das estimativas também são mantidas.

Tabela 12 – Estimativas do modelo linear generalizado gama do IMC com dados imputados pelo PMM (Cenário 1) (n = 1980)

Covariável	Estimativa	Erro padrão	Teste T	Valor-p
Ordenada na origem	20.373	0.542	38.847	<0.001
Escolaridade	0.032	0.029	1.103	0.270
Idade	0.105	0.012	8.961	<0.001
Estado civil (Casado)[+]	0.474	0.215	2.200	0.028
Estado civil (Outro)[+]	0.227	0.436	0.522	0.602
Anos	**0.075**	0.014	5.322	<0.001
***Snack* (Um)**[a]	-0.171	0.239	-0.717	0.473
***Snack* (Dois)**[a]	-0.589	0.273	-2.156	0.031
***Snack* (Três ou mais)**[a]	-0.862	0.325	-2.655	0.008
AIC	10603			
R^2	14.5%			

[+] Categoria referência: solteiro
[a] Categoria referência: Zero *snacks*

5.3.3. Imputação simples por aplicação do índice de propensão (Cenário 1)

Após a estimação do modelo de regressão logística para determinar o índice de propensão para ter a variável escolaridade omissa, na amostra, o mesmo índice foi dividido em quantis. Os resultados da regressão logística foram apresentados anteriormente, na secção de exploração de dados omissos (Tabela 8). Podemos observar no histograma da escolaridade, imputada pelo IP, que a distribuição da variável é semelhante à encontrada nos dados completos, evidenciando o comportamento satisfatório deste método de imputação.

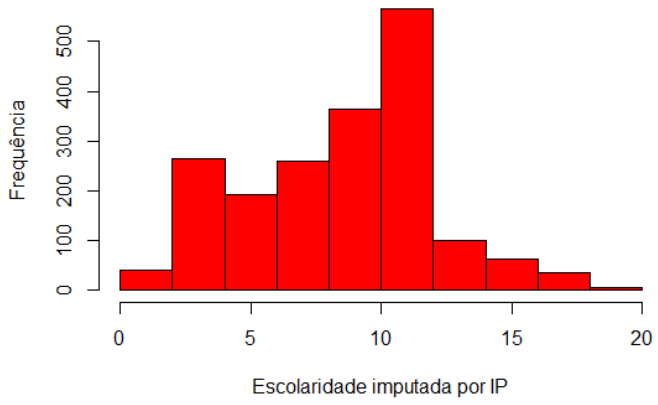


Figura 26 - Histograma da escolaridade imputada por IP (n = 1980)

A tabela 14 mostra os resultados da modelação do IMC, no caso da aplicação da IS por IP à escolaridade.

Tabela 13 – Resultados do modelo de seleção *backwards* no modelo linear generalizado gama do IMC com dados imputados pelo IP (Cenário 1) (n = 1980)

Variável	Modelo	Deviance	valor-p
-	1	47.869	
Origem	2	47.896	0.323
Sexo	3	47.962	0.260
Refeições	4	48.027	0.124

A Tabela 15 apresenta o modelo linear generalizado gama do IMC selecionado, no caso da aplicação da IS por IP à variável escolaridade.

Tabela 14 – Estimativas do modelo linear generalizado gama do IMC com dados imputados pelo IP (Cenário 1) (n = 1980)

Covariável	Estimativa	Erro padrão	Valor T	Valor-p
Ordenada na origem	20.427	0.518	39.399	<0.001
Escolaridade	0.029	0.029	0.985	0.324
Idade	0.105	0.012	8.948	<0.001
Estado civil (Casado)[+]	0.472	0.216	2.188	0.029
Estado civil (Outro)[+]	0.235	0.436	0.540	0.589
Anos	**0.075**	0.014	5.316	<0.001
***Snack* (Um)**[a]	-0.173	0.239	-0.725	0.468
***Snack* (Dois)**[a]	-0.589	0.273	-2.154	0.031
***Snack* (Três ou mais)**[a]	-0.863	0.325	-2.656	0.008
AIC	10603			
R^2	14.38%			

[+] Categoria referência: solteiro
[a] Categoria referência: Zero *snacks*

O modelo obtido através do índice de propensão é muito semelhante àquele encontrado através do PMM. Os valores de coeficientes, significância das variáveis na explicação da variável resposta, AIC e R^2 mantêm-se similares, nos dois casos, incluindo para a variável com imputação - escolaridade.

A Tabela 16 apresenta a distribuição da variável escolaridade, após aplicação dos diferentes processos de imputação.

Tabela 15 – Distribuição da variável escolaridade

	Escolaridade(CC)	Escolaridade(IS^{+})	Escolaridade (IS*)	Escolaridade (IS^{a})
n	1785	1904	1904	1904
Mínimo	1,000	1,000	1,000	1,000
1º Quartil	7,000	7,000	6,803	6,000
Mediana	9,000	9,182	9,000	9,000
Média	9,207	9,182	9,098	9,091
3º Quartil	19,000	12,000	12,000	12,000
Máximo	19,000	19,000	19,000	19,000
NAs	135	0	0	0

CC – Casos Completos; + - IS por substituição pela mediana; * - IS por PMM; a – IS por IP

Pode verificar-se que a distribuição da escolaridade não difere significativamente por tipo de imputação, nem por tratamento de casos completos em comparação com as restantes técnicas. Em seguida, apresentam-se os gráficos *boxplot* da variável escolaridade, após aplicação de cada uma das técnicas para tratar dados omissos.

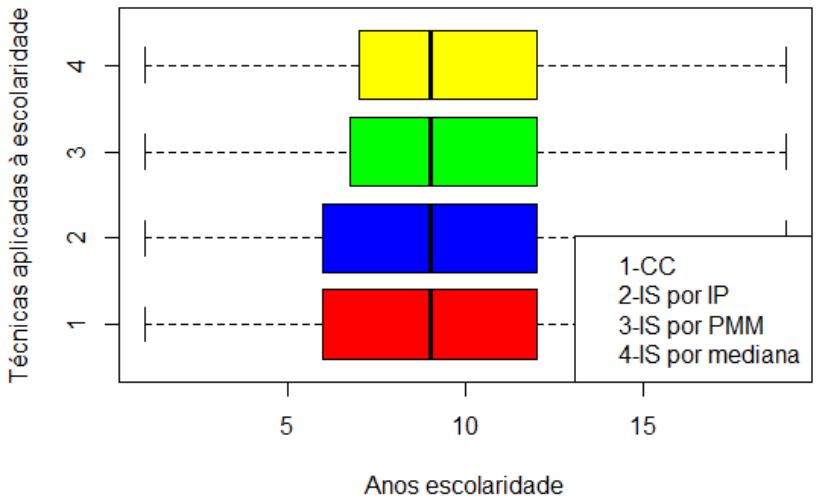


Figura 27 - *Boxplots* da variável escolaridade nos diferentes cenários de tratamento de dados (CC, IS por substituição da mediana, IS por PMM e IS por IP)

A Figura 27 sugere que a distribuição da variável escolaridade é semelhante, independentemente da técnica aplicada. Podemos verificar uma média sensivelmente igual (Tabela 16), nos quatro casos, embora exista maior variabilidade na distribuição da escolaridade nos casos completos e na IS por IP.

5.3.5. Imputação múltipla por *predictive mean matching* (Cenário 1)

Diagnóstico da imputação

A qualidade da imputação é ditada, em parte, pela sua capacidade de prever valores omissos dentro de intervalos realistas, no estudo a ser realizado. Por exemplo, valores à partida impossíveis como frequências negativas ou homens grávidos, não deverão surgir nos dados imputados (57). Assim sendo, foi verificada a plausilabilidade dos dados imputados, através da função **impimpescol.** A Figura 28 ilustra os primeiros resultados da IM para cada uma das cinco de bases imputadas. O sumário da variável escolaridade, em cada uma das cinco bases de dados imputadas, é apresentado no Anexo 9.5.

```
> head(imp$imp$escol)
    1  2  3  4  5
7  12 12 12  7 12
9   7  4  4  4  4
10  6  6  9 14 12
17  6 12  7  6 12
18  1  3  3  4  1
19  8  5 16  4  4
```

Figura 28 - Primeiras linhas dos resultados da IM para cada uma das bases de dados imputadas (Cenário 1, IM - PMM)

Pode ser útil verificarmos a distribuição dos dados antes e depois da imputação. Para isso, usamos a função **stripplot()** obtendo o resultado abaixo (Figura 29). É também possível obter um gráfico stripplot() para toda a base de dados, distinguindo entre valores observados e imputados, o que pode ser particularmente útil quando se imputa mais do que uma variável (Anexo 9.6).

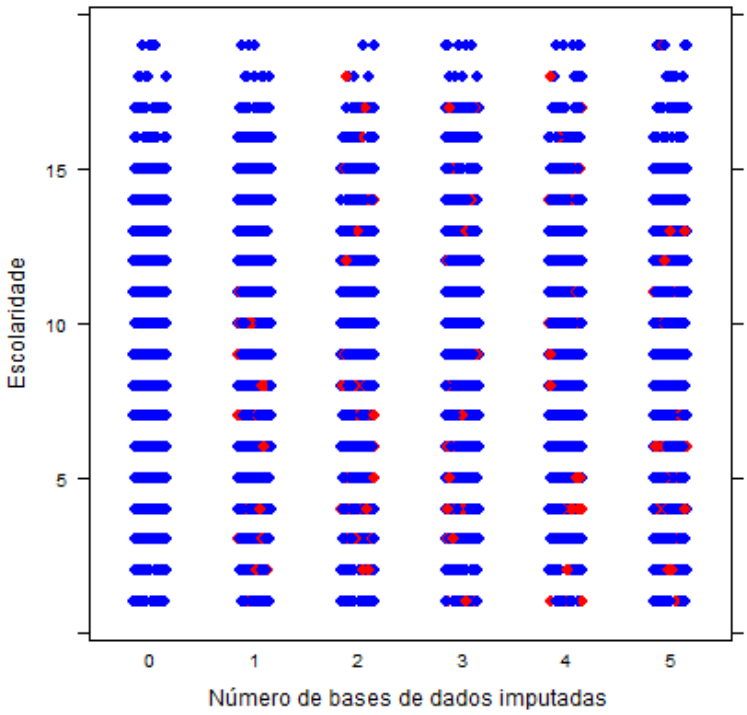


Figura 29 – Distribuição dos dados imputados e reais, na base de dados completa (0) e nas cinco bases de dados com valores imputados (1 a 5) (Cenário 1, IM - PMM)

A Figura 29 representa os valores da escolaridade reais na base de dados, de cor azul, e os valores imputados, de cor vermelha. A primeira coluna corresponde à base de dados com valores omissos para escolaridade. As restantes colunas representam cada uma das novas base de dados geradas pelo processo de imputação. Como o método de imputação usado foi o *predictive mean matching* que só gera valores existentes na base de dados, verificamos que os valores imputados estão sempre dentro dos limites dos valores observados e têm os mesmos intervalos. A figura indica que a distribuição dos valores observados e imputados de escolaridade são semelhantes.

Após a execução da imputação é necessário garantir que os dados imputados são realistas e aceitáveis, perante a realidade clínica estudada pelo investigador. Nesse caso, pode ser útil produzir gráficos da densidade dos valores observados e imputados de todas as variáveis que sofreram imputação (neste caso, escolaridade) para confirmar se as imputações são plausíveis. Diferenças nas densidades dos valores observados e imputados podem indicar problemas maiores que devem ser verificados. A Figura 30 mostra a densidade dos valores observados (a azul) e imputados (a vermelho), nas cinco diferentes bases de dados.

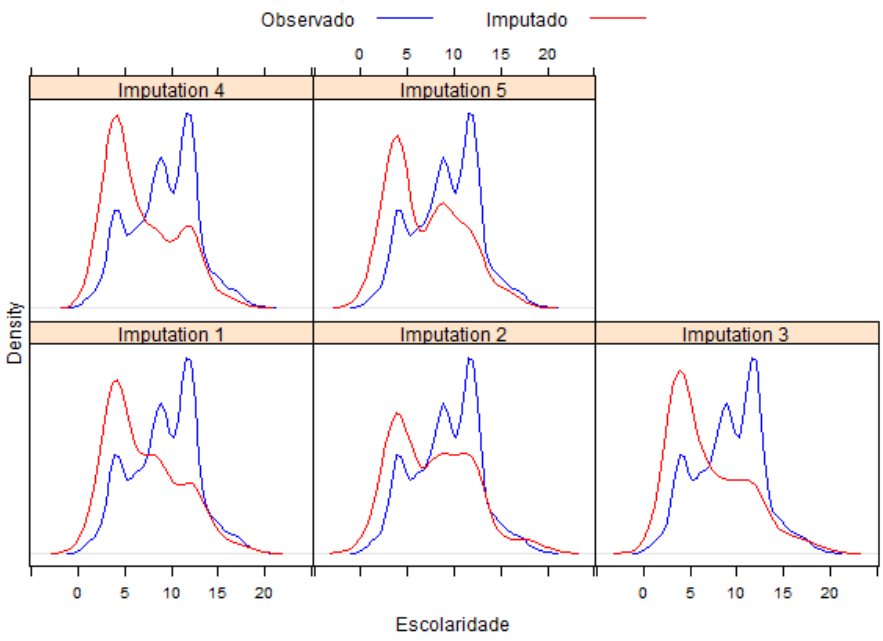


Figura 30 – Gráfico da densidade dos valores observados e imputados da variável escolaridade, por imputação múltipla (Cenário 1, IM - PMM)

Da leitura do gráfico, pode-se verificar que as densidades diferem, sendo os valores imputados de escolaridade tendencialmente mais baixos do que os observados.

Outro método de diagnóstico envolve a comparação da distribuição dos valores observados e imputados, condicionais ao seu índice de propensão (Figura 31). A ideia é que a distribuição condicional deve ser semelhante se o modelo assumido para criar as imputações múltiplas estiver bem ajustado. A figura mostra a escolaridade (dados observados e imputados) *versus* o índice de propensão para ser omisso em escolaridade. Por definição, os valores imputados encontram-se em maior quantidade na metade direita do gráfico. Se o modelo de imputação se ajusta bem, espera-se que para um dado índice de propensão os valores observados e imputados sejam conformes, ou seja, se apresentem distribuídos de forma uniforme. Neste caso, o gráfico sugere estarmos perante essa situação o que nos indica um bom ajustamento da imputação. No entanto, é importante relembrar que este diagnóstico é tão bom quanto o índice de propensão e se não existirem bons preditores dos dados omissos em escolaridade, o diagnóstico gera também conclusões limitadas.

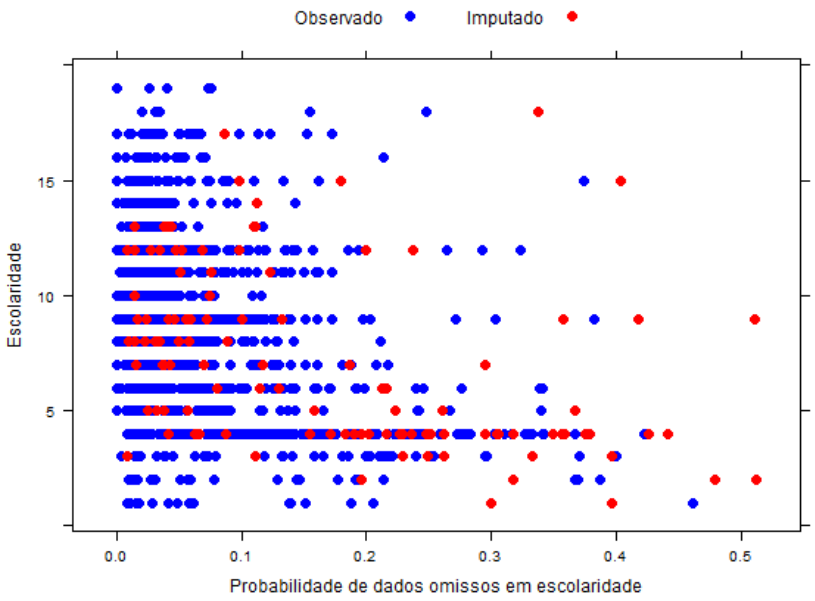


Figura 31 – Valores observados e imputados de escolaridade *versus* índice de propensão (Cenário 1, IM - PMM)

Por fim, produzir o gráfico dos resíduos da regressão da escolaridade em função do índice de propensão, para os valores observados e imputados pode também ser uma ferramenta de avaliação do bom ajustamento da imputação (Figura 32). Um bom indicador deste ajustamento é a sobreposição das duas linhas, o que acontece em grande parte do gráfico abaixo. Assim, este é indicador de que a imputação está bem ajustada.

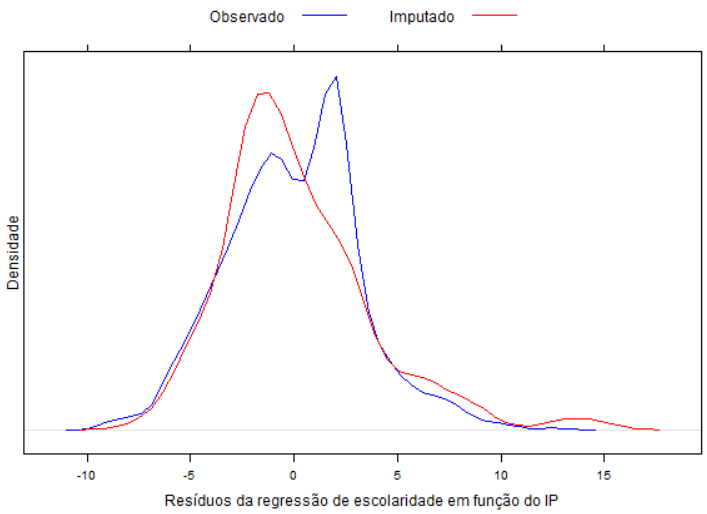


Figura 32 – Resíduos da regressão de escolaridade em função do IP, por valores observados e imputados (Cenário 1, IM - PMM)

Análise dos dados imputados

Para este fim, usou-se a função **with.mids()** que aplica o modelo dos dados completos a cada uma das bases de dados imputadas. O resultado da sua aplicação são cinco análises de bases de dados completas. Para se obter uma análise conjunta utiliza-se a função **pool()**. Esta função calcula a média das estimativas do modelo de dados completos, a variância total das análises repetidas e o aumento relativo na variância, devido à não resposta e a fração de informação omissa (FMI).

A Tabela 17 mostra os resultados dos valores-p do teste Wald, apresentados para cada variável não selecionada, a partir do modelo saturado.

Tabela 16 - Seleção de variáveis para o modelo linear generalizado gama do IMC, através do teste Wald (Cenário 1, IM - PMM)

Variável	Modelo	Valor-p teste Wald
Origem	1	0.311
Sexo	2	0.255
Refeições	3	0.127

A tabela 18 apresenta as estimativas do modelo linear generalizado gama do IMC, após a junção dos resultados dos modelos linares generalizados, referentes às cinco bases de dados imputadas.

Tabela 17 - Estimativas do modelo linear generalizado gama do IMC baseada na análise conjunta das 5 bases de dados imputadas (Cenário 1, IM - PMM)

	Estimativa	EP	Teste T	Valor-p	FMI
Ordenada na origem	20.374	0.531	38.351	<0.001	0.015
Escolaridade	0.032	0.030	1.075	0.282	0.028
Idade	0.105	0.012	8.932	<0.001	0.003
Estado civil (Casado)[+]	0.471	0.215	2.187	0.029	0.001
Estado civil (Outro)[+]	0.225	0.436	0.517	0.605	0.001
Anos	**0.075**	0.014	5.331	<0.001	0.001
***Snack* (Um)**[a]	-0.171	0.239	-0.716	0.474	0.001
***Snack* (Dois)**[a]	-0.590	0.273	-2.158	0.031	0.001
***Snack* (Três ou mais)**[a]	-0.861	0.325	-2.651	0.008	0.001
AIC	10602				
R^2	14.4%				

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

Pode verificar-se que no modelo obtido através da imputação múltipla, a idade ($p < 0.001$), o estado civil ($p = 0.020$), os anos a residir no país ($p < 0.001$) e número de *snacks* consumidos ao longo do dia ($p = 0.031$) têm uma associação significativa com o IMC. A coluna **FMI** contém a fração da informação que está ausente, ou seja, a proporção da variabilidade que é atribuível à incerteza, causada pelos dados omissos (Tabela 18). Todos os valores são inferiores a 0.1, o que de acordo com a tabela previamente descrita e publicada por Molenberghs (3) e tendo em conta o número *m* = 5 de imputações realizadas aponta para uma eficiência relativa elevada (superior a 98%).

Analisou-se também a convergência das iterações, ao fim de cinco imputações, por métodos gráficos (Anexo 9.7). Não parece existir nenhum padrão em particular e as linhas cruzam-se, quase desde o início, o que permite afirmar que estamos perante uma boa convergência.

5.3.6. Imputação múltipla por regressão linear não Bayesiana (Cenário 1)

Diagnóstico da imputação

A imputação por regressão linear não Bayesiana (RLN) gera valores imputados e resultados diferentes da anterior. Pode-se verificar na figura 33, com os valores reais a azul e os valores imputados a vermelho, que as cinco bases de dados com valores imputados seguem um padrão semelhante, tal como seria de esperar. O sumário da variável escolaridade, em cada uma das cinco bases de dados imputadas, no Anexo 9.5.

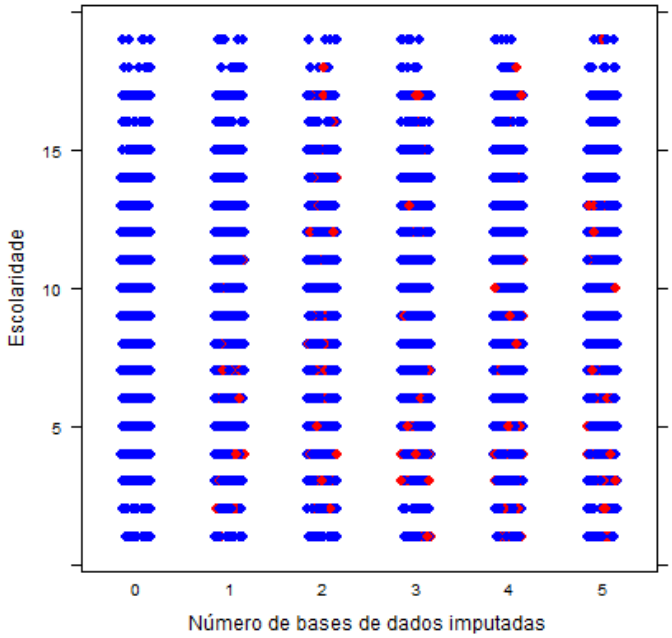


Figura 33 – Distribuição dos dados imputados e reais, na base de dados completa (0) e nas cinco bases de dados com valores imputados (1 a 5) (Cenário 1, IM - RLN)

Analisando a qualidade da imputação, através do gráfico da densidade da variável escolaridade nas cinco bases de dados imputadas, em comparação com a base de dados completos, podemos verificar algumas discrepâncias. Os valores imputados parecem apresentar uma menor densidade em comparação com os valores observados e a sobreposição das linhas é praticamente nula (Figura 34).

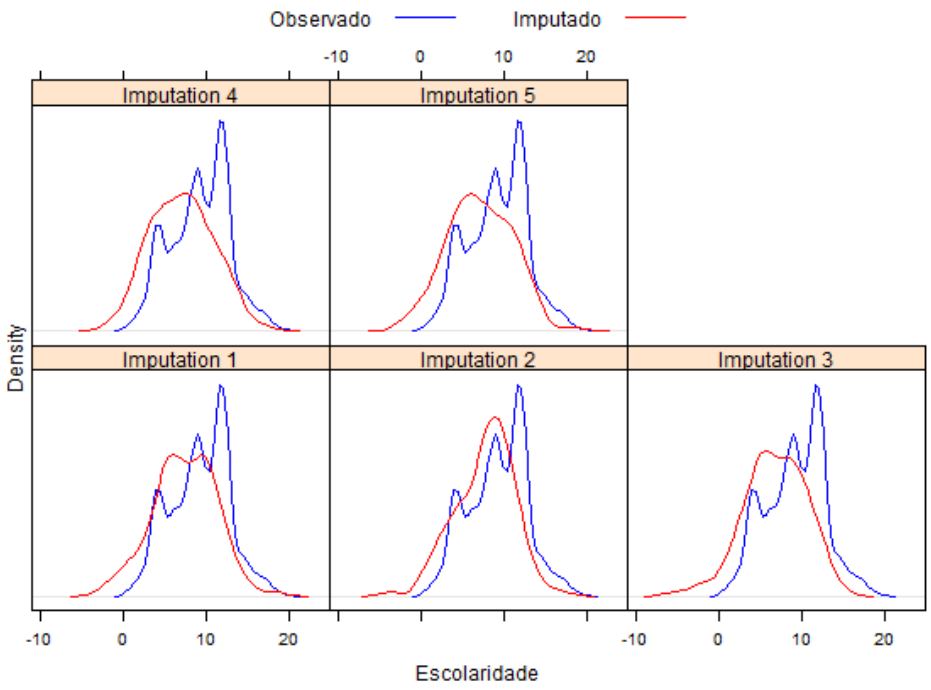


Figura 34 – Gráfico da densidade dos valores observados e imputados da variável escolaridade, por imputação múltipla (Cenário 1, IM - RLN)

Tal como anteriormente, pretendíamos perceber se a distribuição condicional dos dados ao IP é semelhante entre valores omissos e observados.

A Figura 35 mostra a escolaridade (dados observados e imputados) *versus* o índice de propensão para ser omisso em escolaridade. É esperado que para um dado índice de propensão os valores observados e imputados sejam uniformes. Neste caso, parece existir alguma discrepância entre os mesmos, em comparação com o método PMM. Nesta figura, podemos constatar a atribuição de valores negativos à variável escolaridade, nos dados imputados. O método de regressão linear não produz necessariamente valores observados e isso pode ter implicações negativas, com a produção de valores irrealistas.

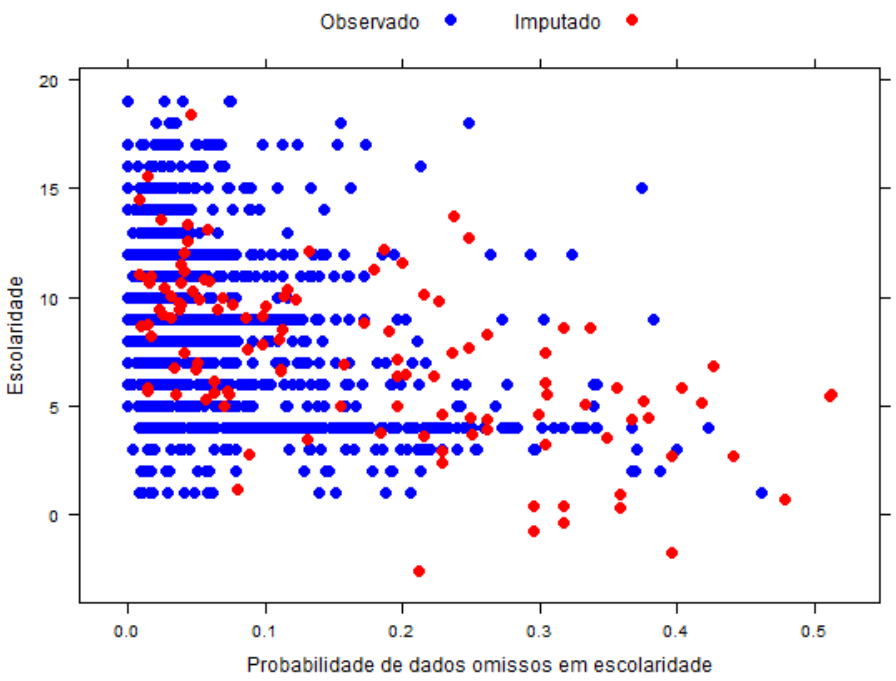


Figura 35 – Valores observados e imputados de escolaridade *versus* índice de propensão (Cenário 1, IM - RLN)

Concluímos a partir da Figura 36, que a densidade dos resíduos da escolaridade, na regressão em função do IP, é semelhante nos valores observados e imputados, visto que as linhas se encontram sensivelmente sobrepostas.

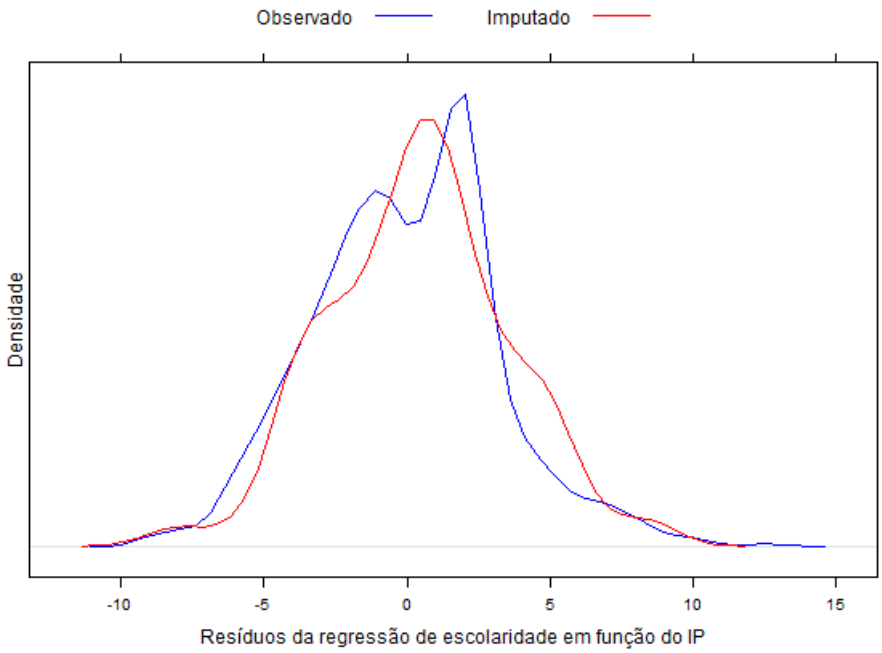


Figura 36– Resíduos da regressão de escolaridade em função do IP, por valores observados e imputados, após imputação múltipla (Cenário 1, IM - RLN)

Análise dos dados imputados

Em seguida, apresenta-se os resultados da seleção de variáveis, através do teste Wald (Tabela 19).

Tabela 18 - Resultados da seleção de variáveis para o modelo linear generalizado gama do IMC, através do teste Wald (Cenário 1, IM - RLN)

Variável	Modelo	Valor-p
Origem	1	0.397
Sexo	2	0.260
Refeições	3	0.126

Os dados imputados deram origem ao modelo linear generalizado gama do IMC apresentado abaixo, na Tabela 20.

Tabela 19 – Estimativas do modelo linear generalizado gama do IMC baseada na análise conjunta das 5 bases de dados imputadas (Cenário 1, IM - RLN)

Variável	Estimativa	Erro padrão	Teste T	Valor-p	FMI
Ordenada na origem	20.450	0.531	38.622	<0.001	0.032
Escolaridade	0.023	0.030	0.757	0.449	0.059
Idade	0.104	0.012	8.852	<0.001	0.007
Estado civil (Casado)[+]	0.474	0.216	2.198	0.028	0.001
Estado civil (Outro)[+]	0.230	0.436	0.483	0.527	0.001
Anos	**0.075**	0.014	5.531	<0.001	0.002
***Snack* (Um)**[a]	-0.171	0.239	-0.717	0.473	0.001
***Snack* (Dois)**[a]	-0.586	0.273	-2.141	0.032	0.002
***Snack* (Três ou mais)**[a]	-0.858	0.325	-2.640	0.008	0.001
AIC	10602				
R^2	14.6%				

[+] Categoria de referência: Solteiro
[a] Categoria de referência: Zero *snacks*

O modelo obtido após a IM por RLN apresenta resultados razoavelmente semelhantes aos anteriores, apesar dos valores negativos gerados. Os anos de residência e a idade apresentam estimativas de coeficiente muito semelhantes aos restantes modelos. A variável escolaridade apresenta o coeficiente mais elevado, do que nas restantes técnicas, embora os EP sejam quase iguais. O R^2 do modelo é igual ao do modelo gerado após IM por PMM e ligeiramente inferior ao dos modelos gerados após IS e CC.

Analisou-se a convergência das iterações, ao fim de cinco imputações, por métodos gráficos (Anexo 9.7). Os gráficos no Anexo 9.7. sugerem uma convergência satisfatória das imputações.

5.3.7. Análise comparativa das técnicas para tratar dados omissos (Cenário 1)

De modo a facilitar a comparação entre os resultados obtidos, criou-se a tabela 21, que apresenta os coeficientes e EP para as variáveis explicativas do IMC, no modelo linear generalizado gama.

Tabela 20 - Efeito de diferentes métodos para lidar com dados omissos nos coeficientes de regressão (β) e erros padrão (EP) das variáveis explicativas induídas no modelo linear generalizado gama do IMC (Cenário 1)

		Imputação simples			Imputação múltipla	
Variáveis	Análise CC	Substituição pela mediana	PMM	IP	PMM	RLN
Ordenada na origem	20.411 (0.381)	20.564 (0.522)	20.373 (0.542)	20.427 (0.518)	20.374 (0.531)	20.450 (0.531)
Escolaridade	0.022 (0.030)	0.018 (0.030)	0.032 (0.029)	0.029 (0.029)	0.032 (0.030)	0.023 (0.030)
Idade	0.104 (0.012)	0.103 (0.011)	0.105 (0.012)	0.105 (0.012)	0.105 (0.012)	0.104 (0.012)
Estado civil (Casado)[+]	0.601 (0.219)	0.478 (0.215)	0.474 (0.215)	0.472 (0.216)	0.471 (0.215)	0.474 (0.216)
Estado civil (Outro)[+]	0.337 (0.451)	0.237 (0.436)	0.227 (0.436)	0.235 (0.436)	0.225 (0.436)	0.230 (0.436)
Anos	0.078 (0.014)	0.075 (0.014)	0.075 (0.014)	0.075 (0.014)	0.075 (0.014)	0.075 (0.014)
Snack (Um)[a]	-0.142 (0.244)	-0.170 (0.239)	-0.171 (0.239)	-0.173 (0.239)	-0.171 (0.239)	-0.171 (0.239)
Snack (Dois)[a]	-0.696 (0.276)	-0.583 (0.239)	-0.589 (0.273)	-0.589 (0.273)	-0.590 (0.273)	-0.586 (0.273)
Snack (Três ou mais)[a]	-0.902 (0.329)	-0.855 (0.325)	-0.862 (0.325)	-0.863 (0.325)	-0.861 (0.325)	-0.858 (0.325)
AIC	9856.6	10603	10603	10603	10602	10602
R^2	15.2%	14.4%	14.5%	14.4%	14.4%	14.6%

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

Os R^2 dos modelos são, no geral, bastante baixos. Pode verificar-se, através da leitura da Tabela 21, que as diferentes técnicas de lidar com dados omissos não implicaram qualquer alteração na direção das estimativas dos coeficientes, sendo esta transversal aos métodos usados. Os R^2 foram semelhantes, nos modelos após imputação, e

ligeiramente inferiores aos encontrados no modelo de casos completos. O AIC foi sensivelmente igual em todos os modelos de imputação (e superior ao modelo de casos completos).. A ordenada na origem apresentou sempre coeficientes semelhantes (embora todos diferentes). Os erros padrão da ordenada apresentavam diferenças ligeiras, sendo a mais discrepante entre o modelo CC (0.381) e os restantes (0.518 – 0.542).

O coeficiente da variável que sofreu imputação, escolaridade, altera com a técnica aplicada (aumenta desde a IS por mediana (0.018) até à IS por PMM e IM por PMM (0.032), sendo 0.022 no modelo CC). Contudo, os EP para a variável são muito semelhantes.

No caso da variável idade e anos, as estimativas dos coeficientes e EP são muito semelhantes entre as técnicas aplicadas. Na variável idade, o coeficiente varia entre 0.103 e 0.105. Na variável anos, todos os coeficientes após imputação são iguais (0.075) e diferem muito pouco do coeficiente obtido na análise CC (0.078). Estas variáveis são também as mais associadas ao IMC, em todos os modelos múltiplos.

A categoria casado da variável estado civil apresenta variação nas estimativas dos coeficientes e EP, de acordo com a técnica implementada. O coeficiente mais alto é o da análise de CC (0.601), sendo os restantes mais baixos, mas muito semelhantes entre si (0.471 – 0.478). Os EP são quase iguais. Na categoria outro da variável estado civil, os coeficientes mais baixos são, mais uma vez, os produzidos pelas imputações, variando entre 0.225 e 0.237. O EP maior é o da análise de casos completos (0.451), sendo os restantes iguais.

Na categoria um da variável *snack*, o coeficiente mais baixo (em módulo) encontrado foi na análise de CC. Os restantes foram semelhantes. Nas categorias dois e três ou mais da mesma variável o coeficiente mais alto (em módulo) encontra-se na análise CC e, mais uma vez, os coeficientes obtidos nas imputações são semelhantes, assim como os EP.

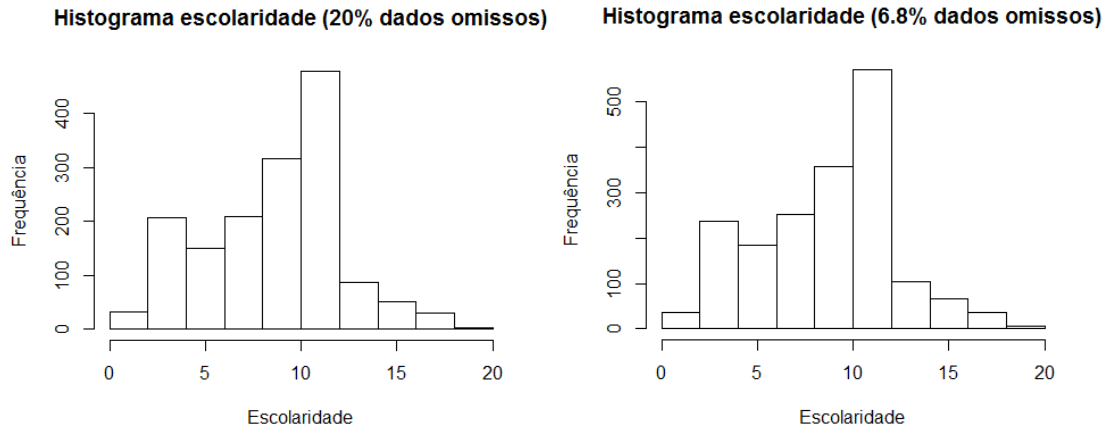
Em suma, verificam-se resultados muito semelhantes, entre técnicas de imputação, independentemente de serem simples ou múltipla. Apesar do diagnóstico da IM por

RLN ter sido menos positivo, os resultados na regressão são muito semelhantes aos restantes. Não há enviesamento na seleção de variáveis, que se mantêm as mesmas, mesmo no modelo de análise CC. No geral, a análise CC produziu EP maiores, mas ainda assim muito próximos dos encontrados nas imputações.

5.4. Cenário 2

5.4.1. Simulação de dados omissos

Procedeu-se à simulação aleatória da falta de 14% dos dados observados de escolaridade, de modo a completar 20% de dados omissos. Apesar destes valores serem obtidos de forma aleatória, estamos perante um mecanismo de omissão MAR, visto que os restantes 6.3% já existiam e foram mantidos. Abaixo apresentam-se os histogramas da variável escolaridade antes e depois da simulação, de modo a verificar-se a distribuição dos dados nas duas situações. Pode constatar-se que a distribuição, após a simulação, se mantém semelhante à inicial, como seria de esperar (Figura 37).



(A) (B)

Figura 37 – Distribuição da variável escolaridade. (A): dados após simulação de 20% de dados omissos, n = 1512 . (B): dados originais, n = 1777.

5.4.2. Análise de casos completos (cenário 2)

A regressão com dados completos utiliza, neste caso, **1512** sujeitos. Os resultados da seleção pelo método *backwards* são mostrados nas Tabelas 22 e 23.

Tabela 21 – Resultados da seleção *backwards* no modelo linear generalizado gama do IMC com dados completos (Cenário 2) (n = 1512)

Variável	Modelo	Deviance	valor-p
-	1	37.136	
Sexo	2	37.142	0.630
Origem	3	37.192	0.177
Refeições	4	37.247	0.156

Tabela 22 – Estimativas do modelo linear generalizado do IMC com dados completos (Cenário 2) (n = 1512)

Covariável	Estimativa	Erro padrão	Teste T	Valor-p
Ordenada na origem	20.295	0.585	34.698	<0.001
Escolaridade	0.018	0.032	0.561	0.547
Idade	0.110	0.013	8.280	<0.001
Estado civil (Casado)[+]	0.597	0.242	2.473	0.014
Estado civil (Outro)[+]	0.335	0.494	0.678	0.498
Anos	0.075	0.016	4.741	< 0.001
***Snack* (Um)**[a]	-0.056	0.269	-0.205	0.838
***Snack* (Dois)**[a]	-0.683	0.300	-2.276	0.023
***Snack* (Três)**[a]	-0.781	0.359	-2.173	0.030
AIC	8414			
R^2	15.5%			

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

Não existem alterações na escolha das variáveis, perante a omissão de 20% dos valores da escolaridade. A principal variável de interesse no modelo de regressão, anos a viver em Portugal, continua associada ao IMC. A intensidade da associação mantém-se, mas com o aumento do erro padrão. Como seria de esperar, com a diminuição do tamanho amostral, o erro padrão das variáveis no modelo aumenta (Tabela 23).

5.4.3. Imputação simples pela substituição da mediana (Cenário 2)

Procedeu-se à imputação dos dados, pela substituição dos valores omissos para escolaridade, pela mediana não condicional. O resultado da distribuição da variável escolaridade, após imputação, é apresentado na Figura 38. Pode-se constatar uma menor dispersão dos dados e maior acumulação em torno da média, tal como seria de esperar após a aplicação desta técnica.

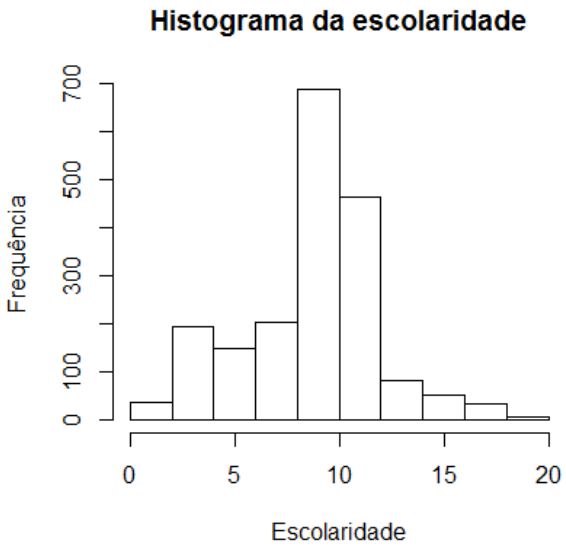


Figura 38– Distribuição da variável escolaridade após IS por mediana (Cenário 2)

Os resultados do processo de seleção de variáveis pelo método *backwards* são mostrados na Tabelas 24.

Tabela 23 – Resultados da seleção *backwards* no modelo linear generalizado gama do IMC com dados imputados pela mediana (Cenário 2) (n = 1896)

Variável	Modelo	Deviance	valor-p
-	1	47.908	
Origem	2	47.927	0.400
Sexo	3	47.962	0.260
Refeições	4	48.027	0.123

Na Tabela 25 encontram-se as estimativas do modelo estimado.

Tabela 24 – Estimativas do modelo linear generalizado do IMC com dados imputados pela mediana (Cenário 2) (n = 1896)

Covariável	Estimativa	Erro padrão	Teste T	valor-p
Ordenada na origem	20.650	0.526	39.249	<0.001
Escolaridade	0.011	0.032	0.348	0.728
Idade	0.103	0.012	8.867	<0.001
Estado civil (Casado)[+]	0.479	0.216	2.223	0.026
Estado civil (Outro)[+]	0.241	0.436	0.552	0.581
Anos	0.074	0.014	5.276	<0.001
***Snack* (Um)**[a]	-0.169	0.239	-0.710	0.478
***Snack* (Dois)**[a]	-0.578	0.239	-2.116	0.034
***Snack* (Três)**[a]	-0.852	0.325	-2.620	0.009
AIC	10604			
R^2	14.4%			

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

A aplicação da IS por substituição da mediana resulta na seleção das mesmas variáveis para o modelo linear generalizado gama final, do que aquelas selecionadas para o modelo gama inicial, no cenário 1, com análise de CC (Tabela 7). Os erros padrão diminuem neste modelo, comparativamente ao anterior, o que seria de esperar visto que voltamos à mesma dimensão amostral. A associação entre anos a viver em Portugal e IMC segue o mesmo padrão, do que os encontrados anteriormente..

5.4.4. Imputação simples pelo *predictive mean matching* (Cenário 2)

Recorreu-se ao método do PMM para imputação dos 20% de dados omissos em escolaridade. Os resultados do processo de seleção de variáveis pelo método backwards são apresentados na Tabela 26. Os resultados das estimativas do modelo selecionado estão na Tabela 27.

Tabela 25 - Resultados da seleção *backwards* no modelo linear generalizado do IMC com dados imputados por PMM (Cenário 2) (n = 1896)

Variável	Modelo	Deviance	valor-p
-	1	47.873	
Origem	2	47.900	0.320
Sexo	3	47.935	0.260
Refeições	4	48.027	0.123

Tabela 26 - Estimativas do modelo linear generalizado do IMC com dados imputados por PMM (Cenário 2) (n = 1896)

Covariável	Estimativa	Erro padrão	Teste T	Valor-p
Ordenada na origem	20.439	0.523	39.060	<0.001
Escolaridade	0.027	0.029	0.931	0.352
Idade	0.105	0.012	8.917	<0.001
Estado civil (Casado)[+]	0.474	0.215	2.200	0.028
Estado civil (Outro)[+]	0.226	0.436	0.518	0.604
Anos	0.075	0.014	5.322	<0.001
***Snack* (Um)**[a]	-0.173	0.239	-0.724	0.469
***Snack* (Dois)**[a]	-0.585	0.273	-2.142	0.032
***Snack* (Três ou mais)**[a]	-0.861	0.325	-2.652	0.008
AIC	10603			
R^2	14.5%			

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

O modelo obtido através do PMM obteve resultados semelhantes aos obtidos através da substituição da mediana, tanto a nível dos coeficientes, como erros padrão e R^2.

5.4.5. Imputação simples pelo índice de propensão (Cenário 2)

O modelo que se segue foi obtido a partir da imputação dos dados omissos pela aplicação do índice de propensão. As tabelas dizem respeito ao processo de seleção *backwards* de variáveis para o modelo linear generalizado gama do IMC, após a imputação mencionada (Tabela 28) e aos resultados do modelo de linear generalizado gama do IMC selecionado (Tabela 29). Recorde-se que o IP é calculado com base no modelo de regressão apresentado na secção 4.2 (Tabela 8).

Tabela 27– Resultados da seleção *backwards* no modelo com dados imputados por IP (Cenário 2) (n = 1896)

Variável	Modelo	Deviance	valor-p
-	1	37.136	
Sexo	2	37.142	0.630
Origem	3	47.962	0.341
Refeições	4	48.027	0.124

Tabela 28– Estimativas do modelo linear generalizado gama do IMC com dados imputados por IP (Cenário 2) (n = 1896)

Covariável	Estimativa	Erro padrão	Teste T	p-value
Ordenada na origem	20.744	0.500	41.544	<0.001
Escolaridade	0.0005	0.028	0.017	0.986
Idade	0.102	0.012	8.829	<0.001
Estado civil (Casado)[+]	0.481	0.216	2.233	0.026
Estado civil (Outro)[+]	0.246	0.436	0.564	0.572
Anos	0.074	0.014	5.253	<0.001
***Snack* (Um)**[a]	-0.168	0.239	-0.703	0.482
***Snack* (Dois)**[a]	-0.575	0.273	-2.102	0.036
***Snack* (Três)**[a]	-0.845	0.325	-2.606	0.009
AIC	10604			
R^2	14.4%			

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

Mais uma vez, mantêm-se as variáveis selecionadas, assim como o R^2 e erros padrão associados a cada variável selecionada no modelo, o que demonstra o bom comportamento das diferentes técnicas de imputação simples aplicadas neste trabalho.

5.4.6. Imputação múltipla por PMM (Cenário 2)

Diagnóstico da imputação

As primeiras linhas do resultado da imputação são mostradas abaixo (Figura 39). É também possível consultar o anexo, que a apresenta a distribuição da escolaridade, nas diferentes bases de dados imputadas (Anexo 9.5).

```
> head(imp$imp$escol)
    1  2  3  4  5
6   7 16 12  9  4
7   9 11  9 12 11
9   4  2  4  7 11
10 10 13  5 12  9
16 11  9  8 10  9
17 12  5 11  9 12
```

Figura 39 - Primeiras linhas dos valores de escolaridade imputados para cada base de dados (Cenário 2)

A Figura 40 representa os valores da escolaridade reais na base de dados, de cor azul, e os valores imputados, de cor vermelha. Pode constatar-se que as distribuições dos valores observados e imputados de escolaridade são semelhantes.

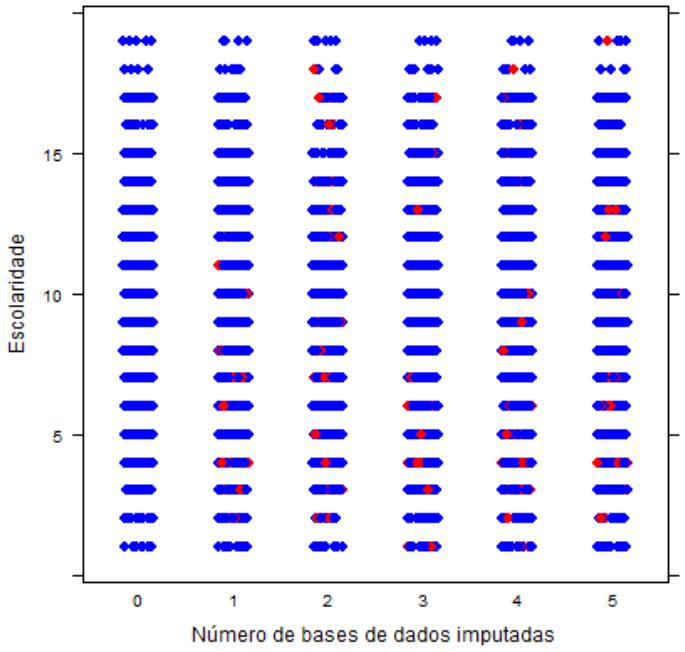


Figura 40 – Distribuição dos dados imputados e reais, na base de dados completa (0) e nas cinco bases de dados (1 a 5) (Cenário 2, IM - PMM)

Para avaliar a plausibilidade das imputações realizadas, produziu-se o gráfico da densidade dos valores observados e imputados, nas cinco bases de dados (Figura 41). A figura revela-nos que a densidade da variável escolaridade é relativamente semelhante, nos dados observados e nos dados omissos. A densidade dos dados imputados aparenta ser mais próxima da dos dados reais, nesta IM por PMM, em comparação com a anterior, feita para apenas 6.8% dos dados de escolaridade (secção 4.3.4.1, Figura 34).

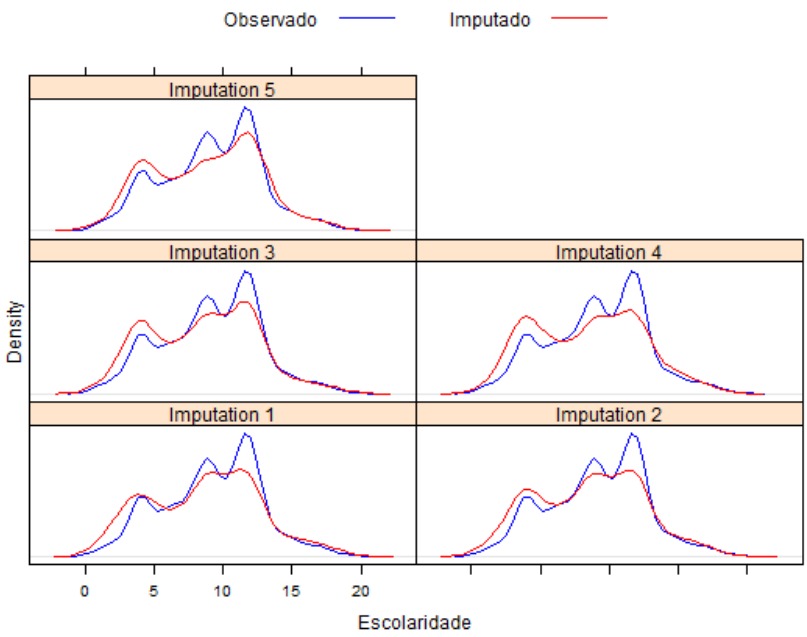


Figura 41– Densidade da escolaridade (Cenário 2, IM - PMM)

Tal como anteriormente, pretendíamos perceber se a distribuição condicional dos dados ao IP é semelhante entre valores omissos e observados.

A Figura 42 mostra a escolaridade (dados observados e imputados) *versus* o índice de propensão para ser omisso em escolaridade. É esperado que para um dado índice de propensão os valores observados e imputados sejam conformes. Parece existir concordância entre os mesmos, tal como foi verificado anteriormente, com a aplicação do mesmo método de IM, para o cenário real 1.

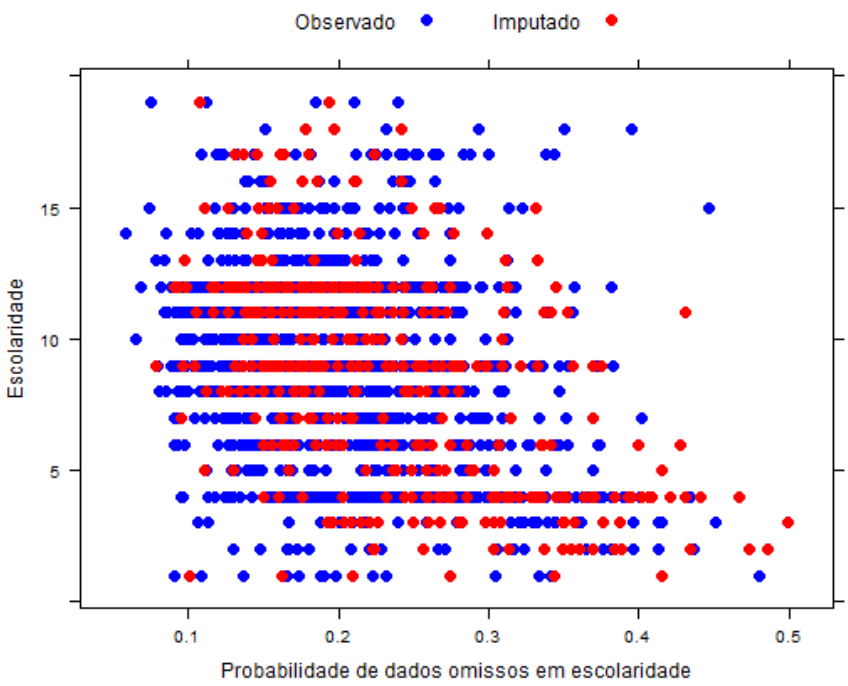


Figura 42 – Probabilidade de dados omissos em escolaridade (Cenário 2, IM – PMM)

A Figura 43 mostra os resíduos da regressão da escolaridade em função do IP, para dados observados e imputados. As linhas encontram-se quase totalmente sobrepostas, revelando uma imputação bem ajustada.

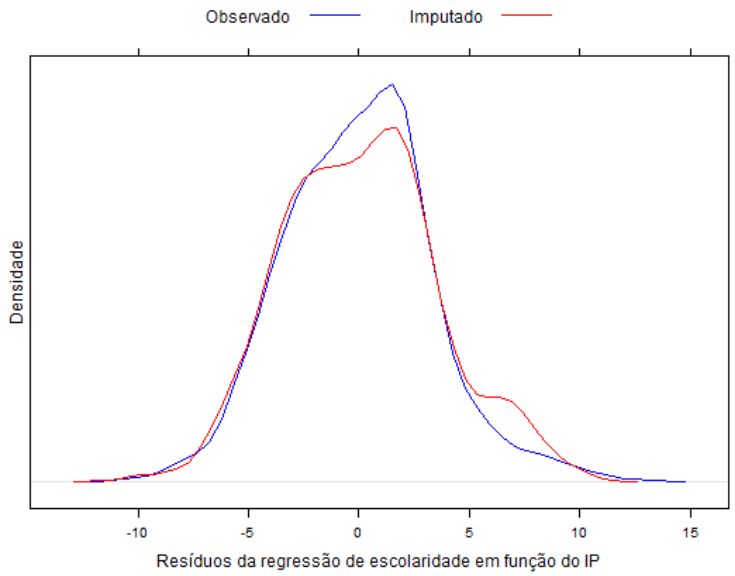


Figura 43 – Resíduos da regressão em função do IP (Cenário 2, IM - PMM)

Análise dos dados imputados

Em seguida, apresenta-se a Tabela 30 com os resultados do teste Wald para a seleção de variáveis para o modelo de linear generalizado gama do IMC.

Tabela 29– Resultados da seleção de variáveis para o modelo linear generalizado gama do IMC, através do teste de Wald (Cenário 2, IM - PMM)

Variável	Modelo	Valor-p do teste Wald
Origem	1	0.397
Sexo	2	0.260
Refeições	3	0.127

A Tabela 31 expõe os resultados do modelo linear generalizado gama final, após a seleção de variáveis pelo teste Wald.

Tabela 30– Estimativas do modelo linear generalizado gama do IMC baseada na análise conjunta das 5 bases de dados imputadas (Cenário 2, IM - PMM)

Variável	Estimativa	EP	Teste T	Valor-p	FMI
Ordenada na origem	20.646	0.540	38.225	<0.001	0.123
Escolaridade	0.020	0.032	0.635	0.530	0.228
Idade	0.101	0.012	8.617	<0.001	0.019
Estado civil (Casado)[+]	0.434	0.220	1.976	0.048	0.018
Estado civil (Outro)[+]	0.214	0.427	0.502	0.615	0.002
Anos	**0.075**	0.014	5.372	<0.001	0.002
***Snack* (Um)**[a]	-0.171	0.238	-0.526	0.599	0.001
***Snack* (Dois)**[a]	-0.582	0.273	-2.037	0.042	0.002
***Snack* (Três)**[a]	-0.857	0.325	-2.333	0.019	0.001
AIC	10603				
R^2	14.6%				

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

No modelo linear generalizado obtido após IM por PMM, para substituir 20% dos dados omissos na variável escolaridade, os resultados obtidos são semelhantes àqueles encontrados no modelo original de casos completos (cenário 1) e nos modelos de IS, tanto quando a variável escolaridade apresenta cerca de 7% de dados omissos, como quando apresenta 20%. Os erros padrão são semelhantes aos encontrados anteriormente e o R^2 médio da regressão é quase igual ao apresentado na regressão após IM por PMM e RLN, quando a variável escolaridade apresentava uma baixa

percentagem de dados omissos (ou seja, no primeiro cenário apresentado). Este valor é ligeiramente inferior ao encontrado nas regressões após aplicação da IS. A associação entre anos a viver em Portugal e IMC permanece igual, tanto a nível da estimativa do coeficiente, quanto do erro padrão associado.

Foi avaliada a convergência das iterações do modelo, por meios gráficos (Anexo 9.7), que mostram uma convergência satisfatória.

5.4.7. Análise comparativa das técnicas para tratar dados omissos (Cenário 2)

A tabela 32 apresenta o efeito dos diferentes métodos para lidar com dados omissos, usados ao longo do cenário 2, nas estimativas dos coeficientes de regressão e erros padrão das variáveis explicativas, selecionadas para o modelo linear generalizado gama do IMC.

Tabela 31– Efeito de diferentes métodos para lidar com dados omissos nos coeficientes de regressão (β) e erros padrão (EP) das variáveis explicativas incluídas no modelo linear generalizado gama do IMC (Cenário 2)

			Imputação simples			**Imputação Múltipla**
Variáveis	Análise CC [b]	Análise CC[c]	Substituição pela mediana	PMM	IP	PMM
Ordenada na origem	20.411 (0.381)	20.295 (0.585)	20.650 (0.526)	20.439 (0.523)	20.744 (0.500)	20.646 (0.540)
Escolaridade	0.022 (0.030)	0.018 (0.032)	0.011 (0.032)	0.027 (0.029)	0.001 (0.028)	0.020 (0.032)
Idade	0.104 (0.012)	0.110 (0.013)	0.103 (0.012)	0.105 (0.012)	0.102 (0.012)	0.101 (0.012)
Estado civil (Casado)[+]	0.601 (0.219)	0.597 (0.242)	0.479 (0.216)	0.474 (0.215)	0.481 (0.216)	0.434 (0.220)
Estado civil (Outro)[+]	0.337 (0.451)	0.335 (0.494)	0.241 (0.436)	0.226 (0.436)	0.246 (0.436)	0.214 (0.427)
Anos	0.078 (0.014)	0.075 (0.016)	0.074 (0.014)	0.075 (0.014)	0.074 (0.014)	0.075 (0.014)
***Snack* (Um)**[a]	-0.142 (0.244)	-0.056 (0.269)	-0.169 (0.239)	-0.173 (0.239)	-0.168 (0.239)	-0.171 (0.238)
***Snack* (Dois)**[a]	-0.696 (0.276)	-0.683 (0.300)	-0.578 (0.239)	-0.585 (0.273)	-0.575 (0.273)	-0.582 (0.273)
***Snack* (Três ou mais)**[a]	-0.902 (0.329)	-0.781 (0.359)	-0.852 (0.325)	-0.861 (0.325)	-0.845 (0.325)	-0.857 (0.325)
AIC	9856.6	8414	10604	10603	10604	10603
R^2	15.2%	15.5%	14.4%	14.5%	14.4%	14.6%

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*
[b] Análise CC "real" (Cenário 1)
[c] Análise CC após simulação (Cenário 2)

A análise CC, no cenário 2, apresenta as maiores discrepâncias, relativamente tanto à análise CC "real" do cenário 1, como às análises após aplicação de outras técnicas para lidar com dados omissos. Pode constatar-se que a omissão de 20% de dados na variável escolaridade não afeta a seleção de variáveis para o modelo. Este é resultado esperado, visto que a eliminação dos dados da variável escolaridade foi aleatória. A análise de CC, no cenário, não altera significativamente a associação entre IMC e a principal variável de interesse anos de residência em Portugal, resultando num coeficiente e num erro padrão semelhantes aos "reais" . De resto, todos os modelos obtidos resultam em estimativas bastante semelhantes da medida desta associação. No entanto, constata-se que no geral, a análise CC no cenário 2, apresenta coeficientes semelhantes e erros padrão acima dos "reais" (cenário 1).

No caso das imputações, verifica-se uma subestimação dos coeficientes, no geral, mais flagrante nas variáveis *snack* e estado civil, comparativamente com as análises CC. Na maioria dos casos, a IM produz EP superiores ou iguais aos encontrados na IS. Ainda assim, a análise CC do cenário 2, que possuía um valor amostral menor que as restantes, produziu os EP maiores, entre todas.

Os coeficientes menos afetados pelas técnicas aplicadas são, mais uma vez, a idade e os anos a viver em Portugal. A variável imputada, escolaridade, apresenta diferenças, dependendo da técnica aplicada. Os valores mais discrepantes detetam-se na análise após imputação por IP (0.001±0.028).

No geral, detetam-se algumas diferenças entre a análise CC do cenário 2 e as restantes análises após imputação. A primeira apresenta erros padrão superiores, assim como algumas alterações a nível dos coeficientes. Não se verificam diferenças significativas nas *performances* de cada uma das técnicas de imputação, sugerindo que a performance de técnicas de IS e IM, neste caso, é similar.

5.5. Cenário 3

5.5.1. Simulação de dados omissos (Cenário 3)

Para que possa ser compreendido o efeito que a existência de valores omissos pode ter na seleção de variáveis, optou-se por simular que uma das variáveis mais associadas ao IMC tivesse 20% de valores omissos. Estes valores omissos foram obtidos de forma aleatória e estamos, por isso, perante dados omissos do tipo MCAR. A vantagem da simulação é que podemos comparar os resultados obtidos a partir da análise de dados completos e das análises com recurso a imputação com os resultados dos dados observados.

Os histogramas abaixo mostram a distribuição da variável idade antes e depois da simulação. Ou seja, no primeiro caso apresentam-se os dados observados (n = 1980). No segundo caso, temos o produto da remoção de 20% das observações da idade, totalizando 1422 observações (Figura 44).

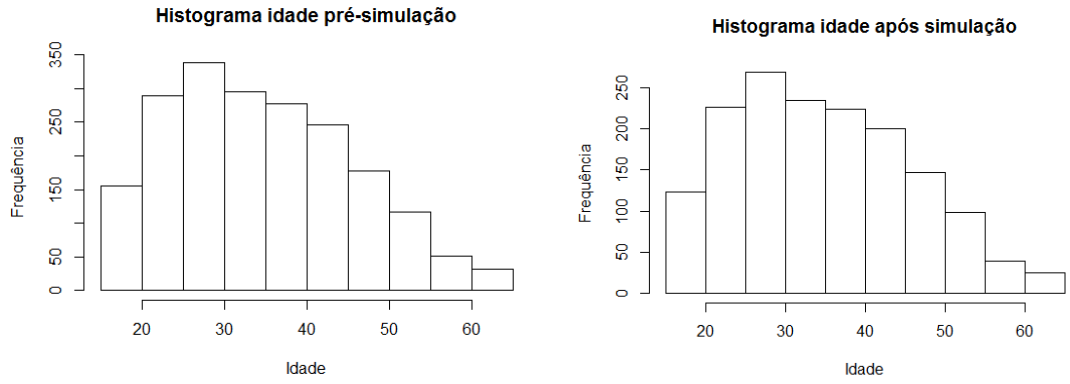


Figura 44 - Histograma da variável idade antes (n = 1980) e depois da simulação(n = 1422)

A distribuição das duas variáveis é também apresenta na Tabela 33. Pode-se constatar através da observação dos histogramas ou da tabela que, tal como seria de esperar, as distribuições são semelhantes.

Tabela 32 - Distribuição da variável idade antes e depois da simulação

	Idade observada	Idade com perda de 20% dos casos
Mínimo	18.0	18.0
1º Quartil	26.0	26.0
Mediana	34.0	34.0
Média	35.2	35.3
3º Quartil	43.0	44.0
Máximo	64.0	64.0
NAs	0	396

5.5.2. Análise de casos completos (Cenário 3)

A Tabela 34 apresenta os resultados do processo de seleção *backwards* no modelo linear generalizado do IMC. Utilizou-se a base de dados completa, após a simulação.

Tabela 33 - Resultados da seleção *backwards* no modelo linear generalizado gama do IMC com dados completos (Cenário 3) (n = 1422)

Variável	Modelo	Deviance	valor-p
-	1	35.396	
Sexo	2	35.401	0.622
Escolaridade	3	35.411	0.559
Refeições	4	35.444	0.275
Origem	5	35.509	0.125

Tabela 34 - Estimativa do modelo linear generalizado gama do IMC com dados completos (Cenário 3) (n = 1422)

Covariável	Estimativa	Erro padrão	Teste T	p-value
Ordenada na origem	20.458	0.436	46.921	<0.001
Idade	0.112	0.013	8.357	<0.001
Estado civil (Casado)[+]	0.587	0.251	2.335	0.020
Estado civil (Outro)[+]	0.300	0.513	0.584	0.560
Anos	**0.069**	0.016	4.241	<0.001
***Snack* (Um)**[a]	0.002	0.279	0.006	0.996
***Snack* (Dois)**[a]	-0.670	0.311	-2.153	0.032
***Snack* (Três ou mais)**[a]	-0.841	0.374	-2.249	0.025
AIC	8481			
R^2	15.6%			

[+] Categoria referência: Solteiro

[a]Categoria referência: Zero *snacks*

O R^2 é sensivelmente o mesmo, tanto neste modelo como nos modelos obtidos anteriormente, com a variável idade completa (Tabela 35). O modelo de regressão com casos completos apresenta estimativas dos coeficientes na mesma direção das verificadas anteriormente, no modelo linear generalizado gama do IMC com casos completos "real" (Tabela 7). Pode constatar-se que não há diferenças na seleção de variáveis no modelo final. Isto vai de encontro ao esperado, visto que a omissão da idade é MCAR. No caso dos anos a viver em Portugal, a associação permanece positiva e significativa. No geral, como seria de esperar devido ao tamanho amostral, todos os erros padrão aumentam (Tabela 35).

5.5.3. Imputação simples pela substituição da mediana (Cenário 3)

Usou-se, novamente, o método de imputação pela subsituição da mediana. Podemos constatar uma alteração na distribuição da variável, existindo uma concentração muito superior em torno da média (Figura 45).

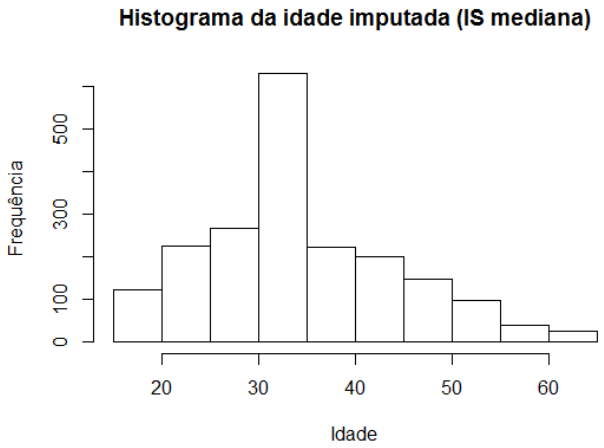


Figura 45 – Histograma da idade imputada por IS por mediana (n = 1980)

Procedeu-se à seleção de variáveis para o modelo de variável resposta IMC (Tabela 36).

Tabela 35 – Resultados da seleção *backwards* no modelo linear generalizado gama do IMC com dados imputados por IS pela substituição da mediana (Cenário 3) (n = 1980)

Variável	Modelo	Deviance	valor-p
-	1	43.556	
Escolaridade	2	48.434	0.693
Sexo	3	48.452	0.414
Origem	4	48.482	0.230
Refeições	5	48.604	0.110

Pode verificar-se, a partir da tabela de estimativas do modelo linear generalizado, que a utilização da IS por substituição da mediana afeta a intensidade das associações entre variáveis (Tabela 37).

Tabela 36 – Estimativas do modelo linear generalizado gama do IMC com dados imputados por IS pela substituição da mediana (Cenário 3) (n = 1980)

Covariável	Estimativa	Erro padrão	Teste T	p-value
Ordenada na origem	20.611	0.415	49.675	<0.001
Idade	0.098	0.013	7.852	<0.001
Estado civil (Casado)[+]	0.767	0.215	3.572	<0.001
Estado civil (Outro)[+]	0.673	0.444	1.515	0.130
Anos	**0.088**	0.014	6.258	< 0.001
***Snack*(Um)**[a]	-0.129	0.245	-0.526	0.599
***Snack*(Dois)**[a]	-0718	0.275	-2.606	0.009
***Snack*(Três ou mais)**[a]	-0.903	0.329	-2.744	0.006
AIC	9872.1			
R^2	14.3%			

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

O modelo de regressão, após a IS por substituição da mediana, apresenta uma seleção de variáveis igual ao modelo de dados completos original (Tabela 6). Os erros padrão são também semelhantes e o R^2 ligeiramente inferior, o que de resto é comum a todos os modelos de imputação. As associações encontradas são, no geral, mais fortes do que aquelas encontradas no modelo original, o que pode estar associado com a diminuição na variabilidade dos dados, consequente da IS. Em particular, a variável

anos apresenta um coeficiente superior ao apresentado no modelo ”real” (cenário 1), mantendo o erro padrão.

5.5.4. Imputação simples pelo *predictive mean matching* (Cenário 3)

A aplicação da técnica PMM para imputar os dados omissos na variável idade revelou-se mais eficaz do que aplicação anterior, com o intuito de imputar os dados omissos na variável escolaridade. O R^2 da regressão que permite prever a idade dos sujeitos é muito superior (44.24%) ao encontrado anteriormente. Os resultados das estimativas do modelo de regressão linear são apresentados na Tabela 38.

Tabela 37 – Estimativas do modelo de regressão linear com variável resposta idade

Covariável	Estimativa	Erro padrão	Teste T	p-value
Ordenada na origem	21.024	1.532	13.725	<0.001
Escolaridade	-0.657	0.062	-10.626	<0.001
Sexo(Masculino)	0.464	0.422	1.100	0.272
Estado civil (Casado)[+]	5.779	0.464	12.448	<0.001
Estado civil (Outro)[+]	9.952	0.944	10.543	< 0.001
Origem (Brasileira)	2.153	0.547	3.939	< 0.001
Refeições (3)	0.287	0.474	0.606	0.544
Anos	0.538	0.034	15.934	< 0.001
***Snack*(Um)**[a]	-0.568	0.518	-1.097	0.273
***Snack*(Dois)**[a]	-0.810	0.607	-1.335	0.182
***Snack*(Três ou mais)**[a]	-2.521	0.720	-3.501	< 0.001
IMC	0.401	0.050	7.964	< 0.001
R^2	44.24%			

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

Pode também constatar-se um valor preditivo razoáveldo modelo de regressão linear, através da observação da Figura 46, com os valores ajustados de idade, a partir do referido modelo, e os valores observados da mesma variável. Embora exista ainda bastante variabilidade entre os valores ajustados e os observados, pode verificar-se uma tendência para que os sujeitos mais velhos sejam classificados como tal e vice-versa.

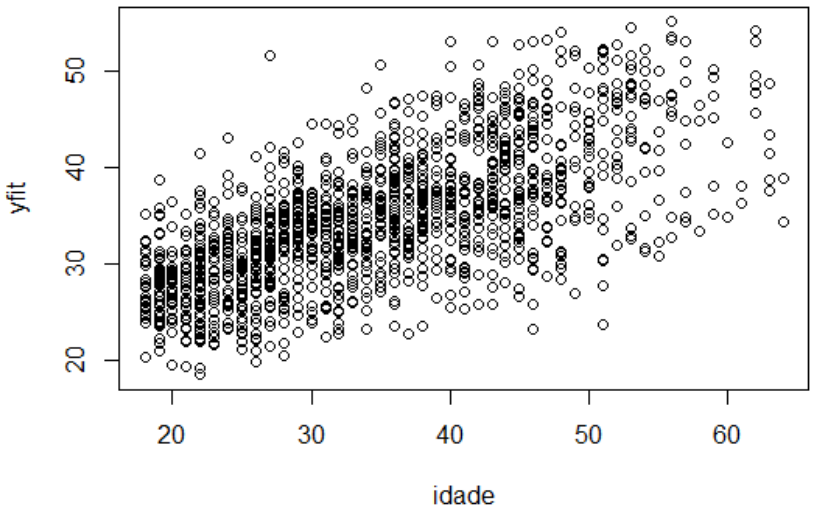


Figura 46 – Valores ajustados da variável idade pela regressão linear

A variável idade, após imputação por PMM, apresenta uma distribuição mais semelhante à original, comparativamente com a distribuição da idade, após IS por mediana, como pode ser verificado pela leitura da Figura 47.

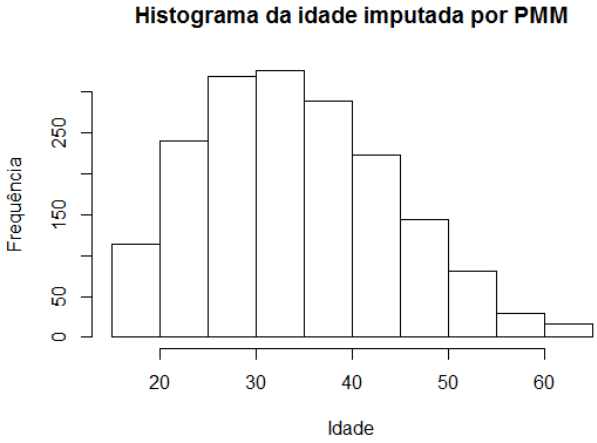


Figura 47– Histograma da idade imputada por IS por PMM (n = 1980)

Procedeu-se, uma vez mais, à seleção de variáveis para o modelo linear generalizado gama do IMC que se apresenta em seguida, na Tabela 39, assim como as estimativas do modelo final, na Tabela 40.

Tabela 38– Resultados da seleção *backwards* no modelo linear generalizado gama do IMC com dados imputados por IS pelo PMM (Cenário 3) (n = 1980)

Variável	Modelo	Deviance	valor-p
-	1	42.478	
Sexo	2	42.486	0.581
Origem	3	42.552	0.112
Escolaridade	4	42.602	0.164
Refeições	5	42.666	0.118
Estado civil	6	42.775	0.100

Tabela 39– Estimativas do modelo linear generalizado gama do IMC com dados imputados por IS pelo PMM (Cenário 3) (n = 1980)

Covariável	Estimativa	Erro padrão	Teste T	Valor-p
Ordenada na origem	20.103	0.389	50.487	<0.001
Idade	0.134	0.011	11.689	<0.001
Anos	**0.060**	0.015	4.112	<0.001
***Snack* (Um)**[a]	-0.140	0.243	-0.577	0.564
***Snack* (Dois)**[a]	-0.640	0.274	-2.335	0.020
***Snack* (Três ou mais)**[a]	-0.809	0.328	-2.469	0.014
AIC	9836.2			
R^2	15.7%			

[a]Categoria referência: Zero *snacks*

O modelo linear generalizado final, com dados imputados por PMM, apresenta divergências com o modelo de casos completos "real" (cenário 1), relativamente às variáveis selecionadas através do processo de seleção *backwards*. A variável estado civil não se encontra no modelo linear generalizado final, ao contrário do que se tinha constatado anteriormente. Anos a viver em Portugal apresenta um coeficiente e erro padrão semelhantes, entre os dois modelos, o que mostra que a associação não difere nos dois cenários (primeiro e último). As variáveis apresentam erros padrão e um R^2 semelhantes aos do modelo de casos completos "real".

5.5.5. Imputação simples pelo índice de propensão (Cenário 3)

Calculou-se, uma vez mais, um índice de propensão para ter dados omissos na variável idade e dividiram-se os sujeitos em grupos por quartis. Apresentam-se em seguida os

resultados do processo de seleção de variáveis para o modelo linear generalizado gama do IMC, após o processo de imputação (Tabela 41).

Tabela 40– Resultados da seleção *backwards* no modelo linear generalizado gama do IMC com dados imputados por IS pelo IP (Cenário 3) (n = 1980)

Variável	Modelo	Deviance	valor-p
-	1	44.131	
Escolaridade	2	44.131	0.886
Sexo	3	44.137	0.638
Origem	4	44.156	0.402
Refeições	5	44.226	0.109

As variáveis selecionadas para o modelo final são as mesmas que no modelo de regressão com dados completos original (Tabela 41).

Tabela 41 - Estimativas do modelo linear generalizado gama do IMC com dados imputados por IS pelo IP (Cenário 3) (n = 1980)

Covariável	Estimativa	Erro padrão	Teste T	p-value
Ordenada na origem	21.423	0.379	56.577	<0.001
Idade	0.067	0.011	6.279	<0.001
Estado civil (Casado)[+]	0.955	0.213	4.492	<0.001
Estado civil (Outro)[+]	0.954	0.444	2.149	0.032
Anos	**0.106**	0.014	7.624	< 0.001
***Snack* (Um)**[a]	-0.170	0.247	-0.691	0.490
***Snack* (Dois)**[a]	-0.746	0.278	-2.683	0.007
***Snack* (Três ou mais)**[a]	-0.944	0.331	-3.000	0.003
AIC	9895			
R^2	13.2%			

[+] Categoria referência: Solteiro
[a]Categoria referência: Zero *snacks*

No modelo linear generalizado, após imputação por IS, pode verificar-se uma maior intensidade das associações das variáveis selecionadas ao IMC, em comparação com o modelo de regressão de dados completos ”real” (cenário 1). Por exemplo, a categoria outro da variável estado civil apresenta um IMC médio significativamente superior à categoria referência solteiro (valor-p = 0.032). Isto não se constata no modelo ”real”,

nem na maioria dos modelos apresentados anteriormente. No geral, todas as variáveis apresentam um valor-p mais significativo neste modelo, do que no modelo ”real”. A variável anos a residir em Portugal apresenta uma estimativa de coeficiente superior indicando que, por cada ano adicional a viver no país, os imigrantes ganham em média 0.106 kg/m^2 (Tabela 42).

5.5.6. Imputação múltipla por PMM (Cenário 3)

Diagnóstico da imputação

Realizou-se uma IM, através do método PMM. Em seguida, apresenta-se as primeiras linhas do resultado desta imputação, na Figura 48. O sumário da variável, obtido por cada base de dados imputada, é apresentado em anexo (Anexo 9.5).

```
> head(imp$imp$idade)
    1  2  3  4  5
8  45 49 36 52 45
12 20 19 19 20 34
15 19 49 23 24 22
20 42 61 41 47 37
21 28 18 32 23 22
25 50 44 43 48 48
```

Figura 48 - Primeiras linhas do resultado da IM por base de dados (Cenário 3, IM - PMM)

A Figura 49 representa os valores da escolaridade reais na base de dados, de cor azul, e os valores imputados, de cor vermelha. Pode constatar-se que as distribuições dos valores observados e imputados de escolaridade são semelhantes.

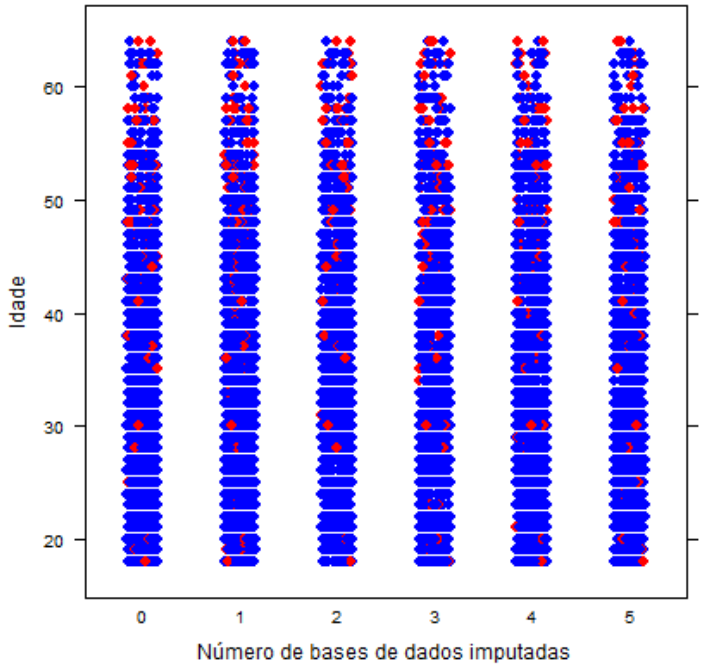


Figura 49 - Distribuição dos dados imputados e reais, na base de dados completa (0) e nas cinco bases de dados com valores imputados (1 a 5) (Cenário 3, IM - PMM)

Para avaliar a plausibilidade das imputações realizadas, produziu-se o gráfico da densidade dos valores observados e imputados, nas cinco bases de dados (Figura 50). A figura revela-nos que a densidade da variável escolaridade é bastante semelhante, nos dados observados e nos dados omissos, com a sobreposição quase total das curvas de gráficos, nas diferentes bases de dados. A densidade dos dados imputados aparenta ser mais próxima da dos dados reais, nesta IM por PMM, do que nas restantes.

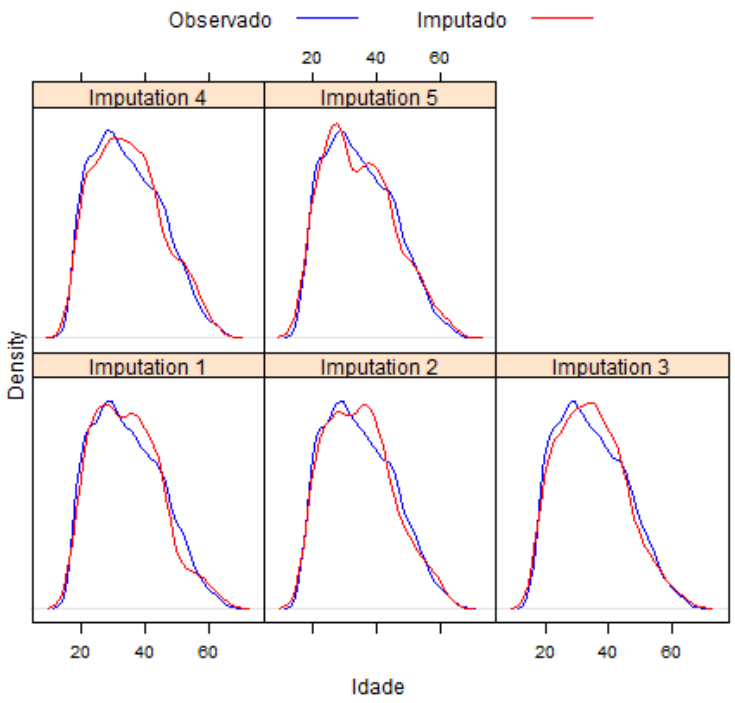


Figura 50– Densidade da idade (Cenário 3, IM - PMM)

A Figura 51 mostra a idade (dados observados e imputados) *versus* o índice de propensão para ser omisso em idade. É esperado que para um dado índice de propensão os valores observados e imputados apresentem uma distribuição unifrome. A observação da figura indica-nos que parece existir concordância entre os mesmos.

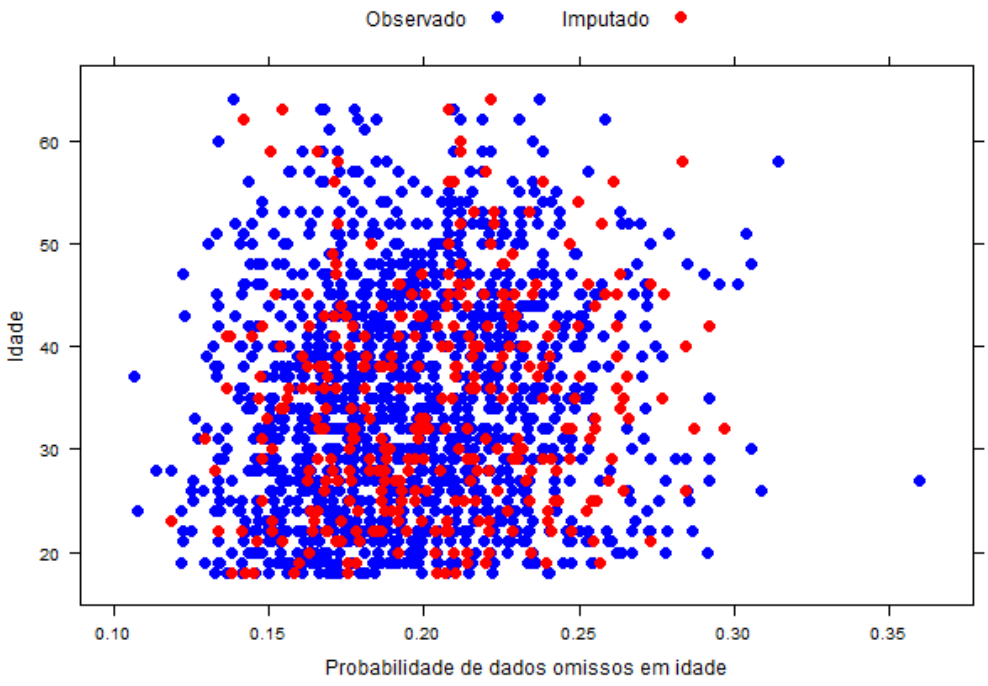


Figura 51– Probabilidade de dados omissos em idade (Cenário 3, IM - PMM)

A Figura 52 mostra os resíduos da regressão da idade em função do IP, para dados observados e imputados. As linhas encontram-se quase totalmente sobrepostas, revelando uma imputação bem ajustada.

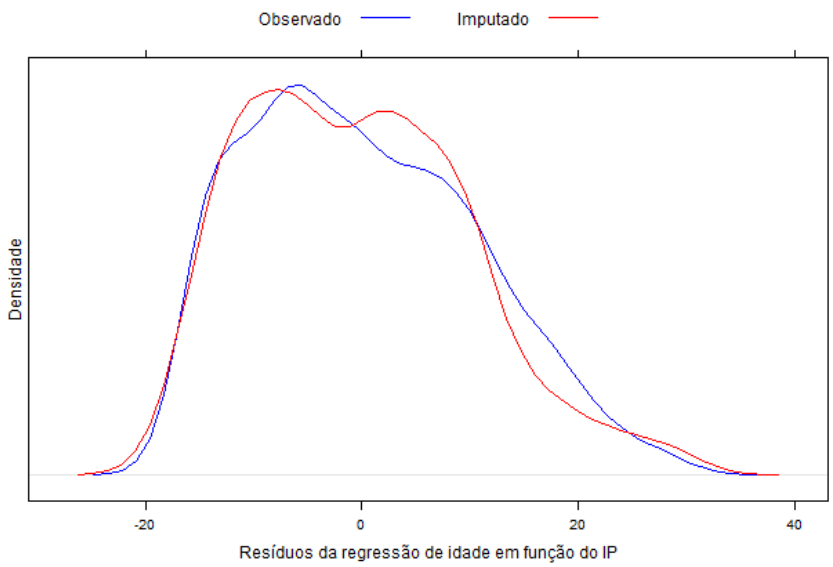


Figura 52 – Resíduos da regressão em função do IP (Cenário 3, IM - PMM)

Análise dos dados imputados

Em seguida, apresenta-se os resultados da seleção de variáveis para o modelo linear generalizado gama do IMC, através do teste Wald (Tabela 43).

Tabela 42 – Resultado da seleção de variáveis para o modelo linear generalizado gama do IMC através do teste Wald (Cenário 3, IM - PMM)

Variável	Modelo	Valor-p
Escolaridade	1	0.318
Sexo	2	0.449
Origem	3	0.289
Refeições	4	0.126

A Tabela 44 expõe os resultados do modelo linear generalizado final, com as variáveis previamente selecionadas.

Tabela 43 – Estimativas do modelo linear generalizado gama do IMC baseada na análise conjunta das 5 bases de dados imputadas (Cenário 3, IM - PMM)

Variáveis	Estimativa	EP	Teste T	Valor-p	FMI
Ordenada na origem	20.764	0.374	55.495	<0.001	0.024
Idade	0.103	0.011	8.920	<0.001	0.039
Estado civil (Casado)[+]	0.479	0.215	2.227	0.026	0.004
Estado civil (Outro)[+]	0.243	0.436	0.557	0.577	0.002
Anos	**0.073**	0.014	5.170	<0.001	0.018
***Snack* (Um)**[a]	-0.181	0.239	-0.759	0.448	0.001
***Snack* (Dois)**[a]	-0.570	0.273	-2.088	0.036	0.002
***Snack* (Três ou mais)**[a]	-0.848	0.325	-2.610	0.009	0.002
AIC	9855.04				
R^2	15.5%				

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*

Podemos constatar que o modelo obtido por IM por PMM apresenta erros padrão muito semelhantes aos encontrados no modelo de regressão com dados completos original (Tabela 7). Os valores-p encontrados são também semelhantes e o R^2 é ligeiramente inferior, embora seja o maior de todos os modelos de IM calculados. A variável anos a viver em Portugal apresenta a mesma associação, com uma estimativa de coeficiente muito semelhante, em ambos os modelos o que demonstra uma boa performance do modelo de IM.

Foi realizada a análise da convergência das iterações, por métodos gráficos, após as cinco imputações (Anexo 9.7). A figura obtida sugere a existência de convergência.

5.5.7. Análise comparativa das técnicas para tratar dados omissos (Cenário 3)

A Tabela 45 apresenta o resumo dos resultados encontrados, nos modelos lineares generalizados gama do IMC, após a aplicação das diferentes técnicas para lidar com dados omissos.

Tabela 44– Efeito de diferentes métodos para lidar com dados omissos nos coeficientes de regressão (β) e erros padrão (EP) das variáveis explicativas induídas no modelo linear generalizado gama do IMC (Cenário 3)

			Imputação simples			Imputação Múltipla
Variáveis	Análise CC[b]	Análise CC[c]	Substituição pela mediana	PMM	IP	PMM
Ordenada na origem	20.411 (0.381)	20.458 (0.436)	20.611 (0.415)	20.103 (0.389)	21.423 (0.379)	20.764 (0.374)
Idade	0.104 (0.012)	0.112 (0.013)	0.098 (0.013)	0.134 (0.011)	0.067 (0.011)	0.103 (0.011)
Estado civil (Casado)[+]	0.601 (0.219)	0.587 (0.251)	0.767 (0.215)	-	0.955 (0.213)	0.479 (0.215)
Estado civil (Outro)[+]	0.337 (0.451)	0.300 (0.513)	0.673 (0.444)	-	0.954 (0.444)	0.243 (0.436)
Anos	0.078 (0.014)	0.069 **(**0.016**)**	0.088 (0.014)	0.060 (0.015)	0.106 (0.014)	0.073 (0.014)
***Snack* (Um)**[a]	-0.142 (0.244)	0.002 (0.279)	-0.129 (0.245)	-0.140 (0.243)	-0.170 (0.247)	-0.181 (0.239)
***Snack* (Dois)**[a]	-0.696 (0.276)	-0.670 (0.311)	-0718 (0.275)	-0.640 (0.274)	-0.746 (0.278)	-0.570 (0.273)
***Snack* (Três ou mais)**[a]	-0.902 (0.329)	-0.841 (0.374)	-0.903 (0.329)	-0.809 (0.328)	-0.944 (0.331)	-0.848 (0.325)
AIC	9856.6	8481	9872.1	9836.2	9895	9855.04
R^2	15.2%	15.6%	14.3%	15.7%	13.2%	15.5%

[+] Categoria referência: Solteiro
[a] Categoria referência: Zero *snacks*
[b] Análise CC "real" (Cenário 1)
[c] Análise CC após simulação (Cenário 3)

Todos os modelos lineares generalizados apresentam um R^2 baixo, embora o modelo com o R^2 mais baixo seja aquele com dados após a aplicação da IS por IP e os modelos com R^2 mais alto os que usou a IS por PMM. No geral, o modelo de análise CC do cenário 3 foi o que mais se distanciou dos restantes. O processo de seleção das variáveis para o modelo linear generalizado final obteve resultados iguais ao modelo "real". Isto não surpreende, visto que o mecanismo de omissão dos dados na idade é MCAR. Os maiores EP são encontrados neste modelo, como seria de esperar, já que a dimensão da amostra é inferior às restantes. Por outro lado, o modelo linear generalizado gama final do IMC com dados após aplicação da IS por PMM, não inclui a variável estado civil que é selecionada em todos os outros modelos calculados. Isto poderá traduzir-se no enviesamento dos resultados e interpretação dos mesmos. Esta técnica subestima, ligeiramente, os coeficientes da variável *snack*. A técnica de IS por

IP sobrestimou, no geral, os coeficientes obtidos, estando quase todos acima do encontrado no modelo "real".

A variável imputada, idade, sofre algumas variação de acordo com a técnica usada para lidar com os dados omissos. As estimativas dos coeficientes mais díspares são obtidos na IS por PMM (0.134) e na IS por IP (0.067), em comparação com a estimativa no modelo linear generalizado após análise de CC do cenário 1 (0.104). No entanto, os valores de EP são muito semelhantes, entre técnicas.

A principal variável de interesse do projeto SAIMI, anos de residência em Portugal, também varia ligeiramente entre técnicas. O valor mais díspar, comparativamente com o "real", foi obtido na regressão após aplicação da IS por IP (0.078 e 0.106, respetivamente). Mais uma vez, os EP são todos relativamente semelhantes, independentemente da técnica.

No geral, as técnicas de IS são menos satisfatórias no Cenário 3. Embora não existam diferenças muito significativas, nas estimativas dos coeficientes e dos EP obtidos, a IS por mediana produz dois dos coeficentes mais díspares, relativamente ao modelo "real", no que toca à idade e anos. A IS por PMM não seleciona as variáveis que seria de esperar, o que pode levar a um enviesamente dos resultados obtidos. A IS por IP sobrestima, no geral, os coeficientes obtidos. Neste cenário, a técnica que obtém um modelo mais próximo do modelo real é a IM por PMM. Em todo o caso, é de salientar que a técnica menos satisfatória para tratar os dados omissos é, claramente, a análise de CC.

6. Discussão

No capítulo de resultados foi apresentada a base de dados completa, a análise bivariada entre as variáveis explicativas e a variável IMC e a análise múltipla, através de um modelo linear generalizado, com o IMC como variável resposta, considerando uma distribuição do tipo gama e função de ligação identidade, para a base de dados completa, sem dados omissos. Em seguida, fez-se uma análise exploratória de dados omissos e optou-se por tratar a variável com maior percentagem dos mesmos (escolaridade), através da IS por substituição da mediana, aplicação do PMM e do IP e através da IM por PMM e RLN. Produziu-se então uma simulação de uma percentagem de dados omissos na variável escolaridade superior aos encontrados, na realidade, no projeto SAIMI, e repetiram-se os procedimentos de IS e IM. Por fim, de modo a compreender o impacto da existência de dados omissos numa variável muito associada à variável resposta, simulou-se a inexistência de 20% dos dados da idade e aplicaram-se as mesmas técnicas de IS e IM.

A análise preliminar dos dados omissos é essencial, em qualquer situação. Estas análises permitem identificar padrões de omissão e possíveis preditores da mesma, auxiliando na escolha de um modelo de imputação correto (58). No presente estudo, procedeu-se a esta análise a partir dos moldes preconizados por Harrell(10), estudando-se o padrão das omissões através da análise de *clusters* e de uma regressão logística. Esta análise permitiu-nos constatar que a omissão de escolaridade seguia um padrão e estava associada à idade, sexo e origem dos sujeitos incluídos nesta amostra. Isto sugere que o padrão de omissão da escolaridade é mais provavelmente MAR e não MCAR, visto depender de variáveis observadas. A abordagem de casos completos requer que os dados omissos sejam MCAR, de modo a produzir resultados não enviesados e generalizáveis para a população em estudo (59).

A variável escolaridade encontra-se associada ao IMC ($r = -0.108$; valor-$p < 0.001$)), na análise bivariada, mas no modelo múltiplo perde a significância estatística. Esta variável não foi selecionada em nenhum dos modelos calculados, pelo método de seleção *backwards*. Optou-se por forçar a sua entrada, de modo a perceber que diferenças produziria, de acordo com o tipo de imputação aplicada. Caso contrário, independentemente da técnica de imputação aplicada, o modelo produziria os

mesmos resultados, visto que a única diferença entre bases de dados com dados imputados era, precisamente, os dados gerados na variável escolaridade.

No primeiro caso, o caso real, a variável escolaridade possuía, originalmente, 6.8% de dados omissos. Após a eliminação das outras observações omissas, nas restantes variáveis, a variável escolaridade passou a apresentar 6.3% de valores omissos, numa base de dados com 1896 observações. Este valor pode ser considerado relativamente baixo e há literatura que sugere que a aplicação de uma técnica de IS pode ser suficiente para tratar o problema (10,53). Todos os modelos produzidos, neste cenário, foram semelhantes. Os modelos imputados apresentam, no geral, um R^2 sensivelmente inferior àquele encontrado no modelo de regressão com casos completos. Os valores dos erros padrão eram também todos sensivelmente semelhantes, revelando uma boa *performance* das técnicas de imputação. Resultados semelhantes foram relatados no passado, em estudos que não encontraram diferenças nos resultados encontrados para análises dos casos completos ou de diferentes técnicas de imputação, quando as variáveis apresentavam valores baixos de dados omissos (abaixo dos 10%) (58).

Optou-se por aplicar três técnicas de IS, em todos os cenários. A substituição pela média ou mediana constitui o tipo de IS mais fácil de aplicar e foi, por esse motivo, incluída neste trabalho. Verificou-se que a sua utilização leva a uma distorção da distribuição da amostra, embora isto nem sempre tivesse implicações vísiveis nos modelos de regressão finais e revelou, também, ter pouco impacto ao nível da seleção de variáveis. O PMM e o IP foram escolhidos por permitirem usar dados da amostra na imputação, de modo a garantir resultados mais fidedignos. A eficácia do PMM depende do coeficiente de explicação obtido no modelo final e que foi, no caso da variável escolaridade, bastante baixo. A seleção de variáveis nos modelos lineares generalizados finais foi, ainda assim, igual à das restantes técnicas e os resultados obtidos semelhantes, no caso da variável escolaridade. O IP é uma técnica ainda pouco usada para efetuar imputação de variáveis contínuas, mas que gerou resultados igualmente interessantes e semelhantes aos encontrados nas restantes técnicas, ainda que se tenham detetado valores-p particularmente baixos no modelo obtido após aplicação desta técnica, na base de dados com 20% de dados omissos em escolaridade.

Aplicaram-se duas técnicas de IM, de modo a comparar os resultados encontrados. Pode constatar-se que a regressão linear não Baeysiana possuía limitações, produzindo resultados pouco realistas, em comparação com o PMM. Este facto não deve ser ignorado, visto que tem de existir coerência nos dados imputados, que devem ser valores realistas da variável com dados omissos. Por outro lado, os resultados obtidos modelo linear generalizado gama do IMC foram silimares em ambos os casos. Após IM por RLN, os erros padrão das estimativas dos coeficientes das variáveis explicativas são ligeiramente superiores àqueles encontrados a partir da PMM, mas esta diferença não parece ter qualquer significado epidemiológico. O que diferencia estes dois métodos é a componente *hot-deck* do PMM, na qual todos os valores imputados são valores da amostra, enquanto que na RLN isto não se passa e são gerados novos valores para imputação. Esta técnica é rápida e eficiente, quando os resíduos do modelo são próximos da normalidade, mas ignora a incerteza associada ao modelo de imputação. No entanto, isto parece ter influenciado pouco os resultados das regressões, sugerindo que qualquer um dos dois métodos pode ser usado (53). Visto que a RLN pode ser usada em amostras grandes (45), é provável que seja esta a razão para ambas as técnicas obterem resultados semelhantes.

A simulação de 20% de dados omissos na variável escolaridade mudou um pouco o cenário encontrado anteriormente. Encontrou-se um modelo linear generalizado gama do IMC com a mesma seleção de variáveis e direção da estimativa de coeficientes, do que os descritos anteriormente, mas com erros padrão superiores nas variáveis selecionadas para este modelo, consequência inevitável da redução da dimensão amostral e que é um dos principais problemas do uso dos casos completos, como estratégia para tratar dados omissos. Tanto a IS como a IM apresentaram resultados semelhantes e que vão de encontro ao encontrado no modelo de regressão com dados completos original e, portanto, o mais próximo da realidade que podemos obter.

Seria de esperar que a IM produzisse estimativas superiores dos coeficientes da regressão, comparativamente com análise de casos completos, mas isto não se verificou. Apenas se registaram alterações ligeiras entre os coeficientes dos modelos. Neste caso, os coeficientes parecem ser insensíveis aos dados omissos e aos diversos modelos de não resposta usados, para lidar com os mesmos. As diferenças no IMC

entre respondentes e não respondentes poderão ser demasiado pequenas para ter impacto nas estimativas (60).

Para se poder perceber se uma variável muito associada ao principal *outcome* pode acarretar um impacto diferente, perante uma elevada percentagem de dados omissos, simulou-se, aleatoriamente, a inexistência de 20% de dados na variável idade. Isto permite estar, sem dúvida, perante um mecanismo de dados omissos MCAR. Neste cenário, não houve consequências a nível da seleção de variáveis, ou da direção das estimativas dos coeficientes, no modelo linear generalizado calculado a partir da base de dados completa. Neste caso, pudemos constatar um comportamento menos satisfatório das técnicas de IS, comparativamente com a IM. A IS por substituição da mediana distorceu os dados, gerando uma distribuição da variável idade menos dispersa e muito centrada em torno da média. No modelo de regressão encontrado verificou-se que esta distorção levou a um aparente aumento da intensidade das associação entre variáveis independentes e IMC. O mesmo se passou na IS por IP, apresentando estas variáveis valores-p mais baixos do que no modelo de regressão com a base de dados original completa. Estes resultados não são surpreendentes, visto que é relatado na literatura a tendência para estas técnicas distorcerem associações e subestimar os erros padrão (35). A IS, através do PMM, deixou uma vez mais a variável estado civil de fora do modelo, embora esta apareça associada com o IMC, no modelo múltiplo da base de dados original completa. Foi, de resto, o modelo de imputação que melhor R^2 apresentou, o que pode ser devido, em parte, a uma boa *performance* do PMM na variável idade. O modelo de regressão linear, com a idade como variável resposta, apresentou um R^2 próximo de 45% e, por isso, a sua capacidade de predizir a idade dos sujeitos que não tinham este dado é melhor, em comparação com o constatado na escolaridade. A IM por PMM mostrou um excelente ajustamento aos dados, gerando valores realistas para a idade, e cuja distribuição se assemelhava bastante à dos dados observados. É de salientar que o facto desta IM ser feita a partir da técnica do PMM, que como se acabou de constatar, apresentou uma boa *performance*, na IS pode ter contribuído para estes resultados. O modelo de regressão produzido a partir da IM é muito semelhante ao encontrado a partir da base de dados completa e original, o que revela o seu comportamento satisfatório. É, de resto,

recomendado o uso da IM, perante uma percentagem elevada (≥ 15%) de dados omissos (10,53).

Os resultados encontrados no presente estudo sugerem que não só a percentagem de dados omissos, como a associação da variável que tem dados omissos ao *outcome* de interesse, têm um papel importante, ao nível da escolha da melhor técnica para lidar com os mesmos. É indicado na literatura que variáveis com uma percentagem intermédia de dados omissos (5 a 15%) poderão ser imputadas, a partir de técnicas de IS, de forma satisfatória (10,53). No presente estudo, verificámos estes resultados. Quando a variável escolaridade apresenta um valor baixo de dados omissos, os resultados do modelo de regressão com base nos casos completos são muito semelhantes àqueles obtidos, após aplicação de técnicas de imputação. Se estamos perante uma percentagem elevada de dados omissos (≥ 15%) é aconselhável recorrer a técnicas de IM, de modo a garantir uma *performance* eficaz da imputação (10,53).

No entanto, é importante salientar que a partir deste único trabalho não é possível tirar conclusões sobre qual o método de imputação mais apropriado para tratar dados omissos, num estudo tranversal, numa variável contínua, pois os resultados obtidos foram bastante semelhantes.

Seria de esperar que os resultados obtidos na regressão com casos completos, após a omissão aleatória de 20% dos dados da idade, fossem semelhantes aos resultados obtidos após IM. Os dados omissos com mecanismo MCAR podem afetar o poder estatístico das análises, mas não irão enviesá-las(2,4,61). É isto que se verifica no presente estudo. O uso da imputação para lidar com dados omissos possui limitações. No caso da imputação simples, acredita-se que produz resultados não enviesados mas sobreestimados na sua precisão (erros padrão demasiado pequenos). No caso da IM, pensa-se que produz resultados não enviesados e com erros padrão corretos. No entanto, deverá existir cautela na utilização desta técnica. A IM assume uma distribuição normal nas variáveis contínuas, que poderá não se verificar. É frequentemente necessário assumir que o padrão de dados omissos é MAR, o que não é possível provar (46). Quando os valores omissos são MNAR, não existe um método uniforme que permita lidar com os mesmos (4). A implementação da IM não deve ser

um processo rotineiro e requer, se possível, o apoio de um profissional especializado(46). No entanto, não existem dúvidas sobre ambos os métodos serem superiores ao método da análise de casos completos. Quando os dados omissos são MCAR, a análise de casos completos pode resultar em associações não enviesadas, mas ineficientes. Quando o tipo de omissão é MAR, o que acontece frequentemente, tem sido extensivamente discutido e demonstrado que uma análise de casos completos é não só ineficiente, mas leva frequentemente a resultados enviesados (6,12). Além disso, estratégias para se lidar com dados omissos podem aumentar o tamanho efetivo do conjunto de dados, contribuindo com o poder necessário para detetar diferenças estatisticamente signficativas(53,59).

Ainda assim, esta noção parece não ser total reconhecida pelos investigadores, já que a maioria dos estudos epidemiológicos ainda usam a análise de casos completos (12). No entanto, regista-se um aumento interessante do uso destas técnicas e um estudo revela que entre 2002 e 2007 a referência a IM, em publicações científicas, duplicou. Existe muito espaço para melhorias, em particular na forma como a informação é relatada, que ainda apresenta muitas falhas (46).

Existem outros métodos para lidar com dados omissos como a estimação de máxima verosimilhança (através da aplicação do algoritmo EM). A estimação de máximo verosimilhança é usada em análises de medidas repetidas ou multinível, quando os preditores e os outcomes são documentados mais do que uma vez, mas no caso de estudos onde os preditores e o outcome são medidos apenas uma vez, a IM é o método preconizado (2,4,61).

Sugere-se que novos estudos incluam mais variáveis com dados omissos e maiores proporções dos mesmos para que se possa verificar o comportamento das técnicas e imputação, nestes diferentes cenários. Para contribuir para a investigação epidemiológica, é essencial surgirem mais trabalhos metodológicos que salientem a importância do tratamento de dados omissos e aplicação de técnicas de imputação, em particular da IM(53). Além disso, os metodologistas deverão focar-se mais na definição de *guidelines* genéricas sobre quando usar técnicas de IM ou IS e por quais optar, de modo a facilitar a aplicação das mesmas(12). A escolha da técnica de

imputação a usar é feita de acordo com a sensibilidade do investigador. No entanto, estes deverão optar por uma técnica de imputação com a qual se sintam confortáveis, tanto metodologicamente como a nível da sua programação. Antes de se fazer esta escolha, os investigadores devem ter em conta as vantagens e desvantagens de cada técnica. Atualmente, com a grande variedade de técnicas de imputação existente, não é adequado a aplicação de uma análise de casos completos, sem que antes se explorem opções de imputação e os seus efeitos (59). Embora se verifique um aumento da sensibilidade científica para a questão dos dados omissos, com um crescente relato da sua existência e, quando possível, dos motivos subjacentes, ainda é pouco comum recorrer à imputação como forma de lidar com dados omissos, parecendo existir um compasso largo de espera entre aquilo que é preconizado pela evidência estatística e aquilo que os investigadores optam por realizar, nos estudos que publicam(62).

7. Conclusão

Embora não se tenha verificado, no presente estudo, diferenças entre as análises obtidas utilizando casos completos e após as imputações, no caso da existência de uma percentagem baixa de dados de omissos em escolaridade, os investigadores devem sempre optar pelo uso de uma técnica de imputação, para lidar com os dados omissos, e garantir que o uso da análise de casos completos não conduz a um viés de seleção, com repercursões a nível das conclusões do estudo(59).

No caso das análises com casos completos, após a exclusão de 20% dos dados da escolaridade e 20% dos dados da idade, a seleção de variáveis foi enviesada e houve perda da poder das análises. Este estudo vai, assim, de encontro a muitos outros e comprova que a análise de casos completos pode levar a conclusões erradas, em estudos epidemiológicos(4,9,12,61).

Neste estudo, não foram encontradas diferenças relevantes no que toca a direção, magnitude e precisão das estimativas dos coeficientes dos modelos lineares generalizados gama do IMC, entre a aplicação de IS e IM, nos dois primeiros cenários (escolaridade omissa com menos de 7%; escolaridade omissa com 20%). No primeiro caso, poder-se-á dever à baixa percentagem de valores omissos (12). No segundo caso,

hipotetiza-se que o facto da variável escolaridade estar pouco associada ao IMC, implicando ser forçada nos modelos finais, pode ter algum peso nos resultados obtidos, alterando pouco os modelos finais (que seriam todos iguais, independente do tipo de imputação usada, caso a variável escolaridade não fosse forçada no modelo final). Aparentemente, a IM não produz resultados superiores, em todas as circunstâncias. Este facto já foi relatado em estudos publicados no passado (12)e pode implicar que sejam necessárias *guidelines* mais específicas para que os investigadores possam saber qual o melhor método de imputação para os seus dados.

A partir deste único estudo, não é possível gerar conclusões gerais acerca dos métodos mais apropriados para se lidar com dados omissos, em modelos múltiplos tranversais. No entanto, o presente estudo conclui que utilizar alguma técnica de imputação é melhor do que ignorar os dados omissos.

8. Bibliografia

1. Goulão B. Excesso de peso nos imigrantes brasileiros e africanos residentes em Lisboa e Setúbal : prevalência e associação com tempo de residência em Portugal. Tese de mestrado. Faculdade de Medicina da Universidade de Lisboa; 2013.

2. Rubin D. Multiple Imputation for Nonresponse in Surveys. Wiley. New York; 1987.

3. Molenberghs G, Kenward MG. Missing Data in Clinical Studies. 1st Edition. Wiley. England; 2007.

4. Moons KGM, Donders R, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006 Oct;59(10):1092–101.

5. Landerman LR, Land KC, Pieper CF. An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. Sociological Methods & Research. 1997 Aug 1;26(1):3–33.

6. Donders ART, Van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006 Oct;59(10):1087–91.

7. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402–6.

8. Nunes LN. Métodos de imputação de dados aplicados na área da saúde. Tese de doutoramento. Faculdade de Medicina da Universidade Federal do Rio Grande do Sul, Brasil; 2007.

9. Rubin DB. Inference and Missing Data. Biometrika. 1976;63(3).

10. Harrell F. Regression Modeling Strategies. Springer New York; 2001.

11. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods. 2002 Jun;7(2):147–77.

12. Van der Heijden G, Donders RT, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. Journal of Clini Epidemiol. 2006 Oct;59(10):1102–9.

13. Mittinty MN, Chacko E. Imputation by Propensity Matching. Survey Research Methods. 2000;4022–8.

14. Allison PD. Multiple Imputation for Missing Data - A Cautionary Tale. Sociol Methods Res. 2000;28(3):301–9.

15. Martins AP. Imputação múltipla - Aplicação prática aos dados do Inquérito ao Emprego. Tese de mestrado. Faculdade de Ciências da Universidade de Lisboa; 2006.

16. World Health Organization. Technical report series 894: obesity: preventing and managing the global epidemic. Geneva: World Health Organization. 2000.

17. National Institutes of Health. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults - The evidence report. 1998.

18. Do Carmo I, Dos Santos O, Camolas J, Vieira J, Carreira M, Medina L, et al. Overweight and obesity in Portugal: national prevalence in 2003-2005. Obes Rev. 2008 Jan;9(1):11–9.

19. Instituto Nacional de Estatística. Indicadores Sociais 2009. Lisboa; 2009 p. 23.

20. Serviço de Estrangeiros e Fronteiras. Relatório de Imigração, Fronteiras e Asilo 2011. Lisboa; 2012.

21. Dias S, Gonçalves A. Migração e Saúde. Migrações. 2007;1:15–26.

22. Satia JA. Dietary acculturation and the nutrition transition. Applied physiology, nutrition, and metabolism. 2010;35:219–23.

23. Delisle H. Findings on dietary patterns in different groups of African origin undergoing nutrition transition. Applied physiology, nutrition, and metabolism. 2010 Apr;35(2):224–8.

24. Colby SE, Morrison S, Haldeman L. What changes when we move? A transnational exploration of dietary acculturation. Ecol Food Nutr. 2009;48(4):327–43.

25. Rosales MV, Jesus VC, Parra S. Crescer Fora de Água. Alto Comissariado Lisboa; 2009. p. 77,91.

26. Gilbert PA, Khokhar S. Changing dietary habits of ethnic groups in Europe and implications for health. Nutr Rev. 2008;66(4):203–15.

27. Arandia G, Nalty C, Sharkey JR, Dean WR. Diet and acculturation among Hispanic/Latino older adults in the United States: a review of literature and recommendations. J Nutr in Geront Geriatr. 2012 Jan;31(1):16–37.

28. Roshania R, Narayan KM, Oza-Frank R. Age at arrival and risk of obesity among US immigrants. Obesity (Silver Spring, Md.). 2008 Dec;16(12):2669–75.

29. Delavari M, Farrelly A, Renzaho A, Swinburn B. Experiences of migration and the determinants of obesity among recent Iranian immigrants in Victoria , Australia. Ethn Health. 2012;37–41.

30. Oza-Frank R, Cunningham S. The weight of US residence among immigrants: a systematic review. Obes Rev. 2010 Apr;11(4):271–80.

31. Dijkshoorn H, Nierkens V, Nicolaou M. Risk groups for overweight and obesity among Turkish and Moroccan migrants in The Netherlands. Public Health. 2008 Jun;122(6):625–30.

32. Toselli S, Galletti L, Pazzaglia S, Gualdi-Russo E. Two-stage study (1990-2002) of North African immigrants in Italy. Homo. 2008 Jan;59(6):439–52.

33. Misra A, Ganda OP. Migration and its impact on adiposity and type 2 diabetes. Nutrition. 2007 Sep;23(9):696–708.

34. Pigott TD. A Review of Methods for Missing Data. Educ Res Eval. 2001 Dec 1;7(4):353–83.

35. Gelman A, Hill J. Missing-data imputation. Data anylsis using regression multilevel/hierarchical models. United States of America: Cambridge; 2007.

36. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol. 2003 Jan;56(1):28–37.

37. Andridge RR, Little RJ. A review of Hot Deck Imputation for Survey Non-response. Int Stat Rev. 2010;78(1):40–64.

38. Little RJ. Missing-Data Adjustments in Large Surveys. Journal Bus Econ Stat. 1988;6(3):287–96.

39. Andreozzi VL. R Function. 2013. Available at valeskaandreozzi.weebly.com/scripts-in-r.htlm

40. Mayer B. Hot Deck Propensity Score Imputation For Missing Values. Science Journal of Medicine and Clinical Trials. 2013;2013(2):1–18.

41. Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Stat Med. 2009;28:1402–14.

42. Rosenbaum P. Observational studies. 2nd Edition. Springer. 2002.

43. Haziza D, Beaumont J-F. On the Construction of Imputation Classes in Surveys. Int Stat Rev. 2007 Apr;75(1):25–43.

44. Buuren S Van. Multiple imputation in practice - Second course. Utrecht; 2011.

45. Groothuis-Oudshoorn K, Buuren S Van. mice : Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67.

46. Sterne J, White I, Carlin J, Spratt M, Royston P, Kenward M, et al. Multiple imputation for missing data in epidemiological studies and clinical research: potencial and pitfalls. BMJ. 2009;338:b2393.

47. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. J R Stat Soc Ser A Stat Soc. 2006 Jul;169(3):1–14.

48. Allison PD. Missing data. Sage University Papers Series on Quantitative Applications in te Social Sciences, 07-136. Thousand Oaks, CA: Sage; 2001.

49. Cohen J, Cohen P. Applied multiple regression and correlation analysis for the behavioral sciences. 2nd Edition. Erlbaum. Hilsdale, NJ; 1985.

50. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat. 2007;61(1):79–90.

51. Buuren S Van, Groothuis-Oudshoorn K, Vink G, Jolani S, Doove L. Multivariate Imputation by Chained Equations. 2013.

52. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001 Dec;6(4):330–51.

53. Maria J, Fachel G. Uso da imputação múltipla de dados faltantes : uma simulação utilizando dados epidemiológicos. Cad Saude Publica. 2009;25(2):268–78.

54. Turkman MAA, Silva GL. Modelos Lineares Generalizados - da teoria à prática. Lisboa: Universidade de Lisboa, Universidade Técnica de Lisboa; 2000. p. 1 – 11.

55. Hruschka DJ. Do Economic Constraints on Food Choice Make People Fat? A Critical Review of Two Hypotheses for the Poverty – Obesity Paradox. Am J Hum Biol. 2012;24:277–85.

56. Dobson A. An introduction to generalized linear models. 2nd Edition. Chapman & Hall/CRC. 2001.

57. Buuren S van, Groothuis-oudshoorn K. MICE : Multivariate Imputation by Chained Equations. Journal of Statistical Software.

58. Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Am J Epidemiol. 2004 Jul 1;160(1):34–45.

59. Bono C, Ried LD, Kimberlin C, Vogel B. Missing data on the Center for Epidemiologic Studies Depression Scale: A comparison of 4 imputation techniques. Res Social Adm Pharm. 2007 Mar;3:1–27.

60. Van Buuren S, Boshuizen H, Knook D. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;18(6):681–94.

61. Little R. Regression With Missing X ' s : A Review. J Am Stat Assoc. 1992;87(420):1227–37.

62. Peugh JL, Enders CK. Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. Review of Educational Research. 2004 Jan 1;74(4):525–56.

9. Anexos

9.1. Programação em R

O R é uma ferramenta para o desenvolvimento de sistemas de apoio à decisão e análise de dados bem como à execução de tarefas mais complexas que envolvam programação, que funciona em Open Source. Isto implica a participação dos utilizadores na construção de *packages* e funções, disponibilizadas para os restantes, e que tornam este sistema comunitário e de acesso livre.

Em particular, para o trabalho desenvolvido nesta dissertação, utilizou-se além das funções incorporadas nos *packages* base, os seguintes *packages* específicos:

- Hmisc (Harell Miscelaneous): permitiu visualizar os padrões de valores omissos, por meios gráficos
- rpart (recursive partition and regression trees): permitiu o estudo da escolaridade pelo método de regressão em árvore
- "VIM" (visualization and imputation of missing values): permitiu a visualização das distribuições marginais das variáveis, tendo em conta dos seus dados omissos, através da função marginplot
- rms (regression modelling strategies): permitiu o calculo do variance inflaction factor (VIF)
- lattice (Lattice graphics): permitiu o diagnóstico das imputações múltiplas
- mice (multivariate imputation by chained equations): *package* específicio para os procedimentos da IM, permitiu efetuar as mesmas, por diferentes métodos, e criar objetos para a regressão ponderada entre as diferentes bases de dados obtidas

As funções específicas usadas, que não se encontravam em nenhum dos *packages* anteriores foram:

- glmfunc.r: função criada por Valeska Andreozzi, permitiu o calculo "automático" dos Odds Ratios das regressões logísticas

- funcimput.r: função criada por Valeska Andreozzi, permitiu a imputação pelos métodos predictive mean matching e índice de propensão

Existem inúmeros documentos de apoio à utilização dos packages mencionados, disponíveis *online*. De salientar, em particular, o material de apoio ao *package* mice que vai além da descrição das funções usadas no mesmo, no artigo publicado por Buuren, na revista *Journal of Statistics Software*(45). O artigo é extremamente elucidativo, relativamente às potencialidades do mice e à forma como este *package* pode ser complementado por outros, tendo permitido um trabalho de diagnóstico e ponderação mais interessante e aprofundado.

9.1.1. Manipulação dos dados

A primeira secção recodifica variáveis, para permitir diferentes análises exploratórias. Certas variáveis, pela sua distribuição, faziam mais sentido recodificadas em categorias. Nem todas as variáveis recodificadas foram, posteriormente, usadas nessa forma nas restantes análises do trabalho.

```
library(foreign)

dados<-read.spss("imc.sav", use.value.labels =
FALSE,to.data.frame=TRUE) #ler dados spss, sem label

names(dados)<-
c("sx","idade","estcivil","escol","natur","anos","imc","refeicoes","sn
ack","origem")#mudar os nomes das variaveis

dados$sx <- factor (dados$sx, labels=c("feminino","masculino"))

dados$origem <- factor
(dados$origem,labels=c("africana","brasileira"))

dados$natur <- factor (dados$natur, labels=c("Angola","Cabo
Verde","Guiné","Moçambique","S Tomé","Outros","Brasil"))

summary(dados)

library(Hmisc)

dados <- dados[dados[,5]!="Outros",]

summary(dados)

dim(dados)

#categorizar refeicoes

dados$refeicoescat<-cut2(dados$refeicoes,c(2.5))
```

```
dados$refeicoescat<-factor(dados$refeicoescat,labels=c("Menos de
3","3"))

summary(dados$refeicoescat)

#categorizar estado civil

dados$estcivilr<-cut2(dados$estcivil,c(2,3))

dados$estcivilr<-
factor(dados$estcivilr,labels=c("solteiro","casado","outro"))

summary(dados$estcivilr)

#categorizar snack

hist(dados$snack)

dados$snackr<-cut2(dados$snack,c(1,2,3))

dados$snackr<-factor(dados$snackr,labels=c("Zero","Um","Dois","Mais de
três"))

summary(dados$snackr)

#categorizar imc

library(Hmisc)

dados$imccat<-cut2(dados$imc,c(18.5,24.9,29.9))

dados$imccat<-factor(dados$imccat,
labels=c("Magreza","Eutrofia","PO","Obesidade"))

summary(dados$imccat)

#categorizar idade

dados$idadecat<-cut2(dados$idade,c(25,35,45,55))

dados$idadecat<-factor(dados$idadecat)

summary(dados$idadecat)

#categorizar anos de residencia

dados$anoscat<-cut2(dados$anos,c(5,10,15))

dados$anoscat<-factor(dados$anoscat,labels=c("0 a 4","5 a 9","10 a
14", "15 ou mais"))

summary(dados$anoscat)
```

9.1.2. Exploração de dados omissos

A exploração de dados omissos seguiu os passos preconizados por Harrell (10) e baseou-se, essencialmente, em ferramentas gráficas de grande interesse. A análise

múltipla realizada consistiu numa regressão logística, sem qualquer seleção de variáveis posterior, visto o objetivo ser apenas encontrar padrões nos dados omissos.

```
#Exploração dos dados omissos

library(Hmisc)

na.patterns<-naclus(dados)

library(rpart)

who.na<-
rpart(is.na(escol)~sx+idade+estcivilr+origem+anos+imc+refeicoescat+sna
ckr, data=dados)

naplot(na.patterns)

na.patterns

naplot(na.patterns,"na per var")

plot(na.patterns)

plot(who.na); text(who.na)

plot(summary(is.na(escol) ~ sx+estcivilr+origem+refeicoescat+snackr,
data=dados))

m<-glm(is.na(escol) ~
sx+idade+estcivilr+origem+anos+imc+refeicoescat+snackr,
family=binomial, data=dados)

summary(m)

1 - (1884/1895) * (768.46/889.24)

#gráficos VIM

library("VIM")

marginplot(dados[, c("escol", "snack")], ylab="Snack",
xlab="Escolaridade", col = c("black","grey"), cex = 1.2,cex.lab = 1.2,
cex.numbers = 1.3, pch = 19)



marginplot(dados[, c("escol", "estcivil")], ylab="Estado civil",
xlab="Escolaridade",col = c("black","grey"), cex = 1.2,cex.lab = 1.2,
cex.numbers = 1.3, pch = 19)



marginplot(dados[, c("escol", "anos")],ylab="Anos de residência",
xlab="Escolaridade", col = c("black","grey"), cex = 1.2,cex.lab = 1.2,
cex.numbers = 1.3, pch = 19)
```

```
marginplot(dados[, c("escol", "refeicoes")],ylab="Refeições",
xlab="Escolaridade", col = c("black","grey"), cex = 1.2,cex.lab = 1.2,
cex.numbers = 1.3, pch = 19)
```

9.1.3. Criação da base de dados completa

Para proceder a análise de uma variável com dados omissos, optou-se por eliminar os valores omissos das restantes. Uma das bases de dados criadas tinha, portanto, valores omissos apenas na escolaridade, enquanto que outra – a base de dados completa – não tinha qualquer valor omisso.

```
retira<-complete.cases(dados[,-4])#ficar sem dados omissos, exceto os
da coluna 4 (escolaridade)

dados<-dados[retira,]

dadoscompletos<-dados[complete.cases(dados),] #dados completos, sem
omissos

summary(dados)

summary(dadoscompletos)

dim(dadoscompletos)

dim(dados)
```

9.1.4. Regressão múltipla com base de dados completa (Cenário 1)

A regressão múltipla com base de dados completa usou apenas os sujeitos que apresentavam todos os valores de cada variável. Optou-se por um método de seleção *backwards*.

```
#Regressão

#método backwards

mod<-
glm(imc~sx+idade+estcivilr+escol+origem+anos+refeicoescat+snackr,famil
y=Gamma(link=identity), data=dadoscompletos)

drop1(mod,test="Chisq")

mod<-
glm(imc~escol+idade+estcivilr+origem+anos+refeicoescat+snackr,family=G
amma(link=identity), data=dadoscompletos)

drop1(mod,test="Chisq")

mod<-
glm(imc~idade+estcivilr+origem+anos+refeicoescat+snackr,family=Gamma(l
ink=identity), data=dadoscompletos)
```

```
drop1(mod,test="Chisq")

mod<-
glm(imc~idade+estcivilr+anos+refeicoescat+snackr,family=Gamma(link=ide
ntity), data=dadoscompletos)

drop1(mod,test="Chisq")

mod<-
glm(imc~escol+idade+estcivilr+anos+snackr,family=Gamma(link=identity),
data=dadoscompletos)

drop1(mod,test="Chisq")

summary(mod)

library(rms)

vif(mod)

#calculo coeficiente de determinação

1 - (1768/1776)*(43246/50745)

#residuos

valajust<-mod$fitted.values

res<-rstandard(mod,type="deviance")

#grafico residuos de deviance

plot(valajust,res,col="lightblue",xlab="valores ajustados",

ylab="resíduos deviance padronizados")

lines(lowess(valajust,res),col="blue")

abline(h=0,lty=2)

#relação linear das variaveis

plot(dados$idade,res,col="indianred1",xlab="idade",

     ylab="resíduos deviance padronizados")

lines(lowess(dados$idade,res),col="red")

abline(h=0,lty=2)

plot(dados$anos,res,col="turquoise",xlab="anos em Portugal",

     ylab="resíduos deviance padronizados")

summary(dados$anos)

lines(lowess(dados$anos,res),col="blue")

abline(h=0,lty=2)
```

```
plot(escol,res,col="maroon",xlab="escolaridade",

     ylab="resíduos deviance padronizados")

lines(lowess(escol,res),col="maroon4")

abline(h=0,lty=2)

##########pontos influentes

h<-hatvalues(mod)

p<-dim(model.matrix(mod))[[2]] # n° parametros

n<-dim(model.matrix(mod))[[1]] #n° observações

plot(h/(p/n),ylab="Leverage h/(p/n)",xlab="índice",
col="lightslateblue")

abline(h=2,lty=2)

plot(h,res, col="lightsalmon",xlab="leverage",ylab="resíduos deviance
padronizados")

abline(h=c(-2,2),lty=2)

abline(v=c(2,3)*mean(h),lty=2)

#distância de Cook

#library(car)

plot(cooks.distance(mod), col="magenta", ylab="Distância de Cook")
```

9.1.5. Imputação simples por mediana (Cenário 1)

A imputação simples por mediana foi feita através da subsituição dos valores omissos pela mediana dos mesmos.

```
#IS mediana

dadosis<-dados #dados para IS - Imputação Simples

dadosis$escol[is.na(dadosis$escol)]<-median(dadosis$escol,na.rm=T) #
Substituição dos omissos na var escolaridade, pela mediana da
escolaridade

dim(dados)

dim(dadoscompletos)

dim(dadosis)

summary(dadosis)

hist(dadosis$escol,col=2)
```

9.1.6. Imputação simples por PMM (Cenário 1)

Foi desenvolvido um *script* específico para este fim por Andreozzi (39). O mesmo script foi, depois, adaptado para se fazer imputação na variável idade.

```
#=============  Método de Imputação: Predictive Mean

#  sintaxe:

#       reg = modelo de regressão para a variável que tem missing utilizando um dara frame com os dados completos

#       dados = data.frame que contém os dados que tem missing na variável que será feita a imputação

impute.predmean<-function(reg,dados){

  #valores observados

  yobs<-reg$model[,1]

  newdata<-
dados[which(is.na(dados$escol)),attr(reg$terms,"term.labels")]

  #função que calcula a distância

dist<-function(x,y){abs(x-y)}

#valores ajustados

  yfit<-reg$fitted

  #valores preditos

  ypredito<-predict(reg,newdata=newdata)

  #calcula a distancia entre cada yfit e ypredito

  d<-outer(yfit,ypredito,FUN="dist")

  #vetor com o valor mínimo da distância para cada ypredito

  minimo<-apply(d,2,min)

  #função que descobre a posição no yfit que se encontra valor mínimo da distância

  findmin<-function(x,min){x==min}

  #vetor com resultado da função findmin
```

```
  posicao<-t(apply(d,1,findmin,min=minimo))

  #caso tenha mais que um valor ajustado cuja distância seja igual ao
valor mínimo, será feito uma amostra aleatória

impute<-
function(pos,y){ifelse(sum(pos)==1,y[pos],sample(y[pos],size=1))}

ynew<-apply(posicao,2,impute,y=yobs)

  return(data.frame(ynew,newdata))

  }
```

9.1.7. Imputação por índice de propensão (Cenário 1)

Foi desenvolvido um *script* específico para este fim por Andreozzi. O mesmo script foi, depois, adaptado para se fazer imputação na variável idade.

```
require(Hmisc)

propscore<-function(mod,varmis,g=4){

  score<-mod$fitted

  grupo<-cut2(score,cuts=quantile(score,probs = seq(0, 1, by=1/g)))

  bbs<-function(x){sample(x[!is.na(x)],replace=T)} #bayesian bootstrap
sample

z<-aggregate(varmis,list(grupo),bbs)

    varmisimpute<-varmis

for (i in 1:g){

    ymis<-
sample(z[,2][[i]],size=sum(is.na(varmis[which(grupo==levels(grupo)[i])
])),replace=T)

length(ymis)

    linha<-cbind(indice<-which(is.na(varmis)), grupo[indice])

    varmisimpute[linha[linha[,2]==i,1]]<-ymis

  }

  varmisimpute

}
```

9.1.8. Imputação múltipla PMM (Cenário 1)

A imputação múltipla pelo PMM, no cenário 1, inicia-se com a escolha dos preditores para o modelo. A função mice, sem nenhuma especificação adicional, considera que todas as variáveis da base de dados são preditoras da variável a ser imputada. O método PMM é *default*, não tendo de ser especificado. O número de imputações a ser feitas, neste caso cinco, também é *default* da função mice.

```
library(mice)
set.seed(145)
imp<-mice(dados,print=FALSE)
pred<-imp$predictorMatrix
pred
pred[,"refeicoes"]<-0
pred[,"estcivil"]<-0
pred[,"natur"]<-0
pred[,"anoscat"]<-0
pred[,"snack"]<-0
pred[,"idadecat"]<-0
pred
```

O código plot(imp) gera a convergência das iterações para todas as variáveis imputadas (neste caso, escolaridade). Pode ser muito útil para avaliar a qualidade da convergência, tal como se pode verificar nos gráficos em anexo.

```
plot(imp)
imp$imp$escol
head(imp$imp$escol)
```

A partir daqui é feito o diagnóstico do modelo de imputação, por meios gráficos.

```
#densidade dos valores observados e imputados da escol
library(lattice)
long <- complete(imp,"long")
levels(long$.imp) <- paste("Imputation",1:5)
long <- cbind(long, escol.na=is.na(imp$data$escol))
```

```
densityplot(~escol|.imp, data=long, group=escol.na, plot.points=FALSE,
ref=TRUE, xlab="Escolaridade",scales=list(y=list(draw=F)),

par.settings=simpleTheme(col.line=rep(c("blue","red"))),

auto.key = list(columns=2,text=c("Observado","Imputado")))
```

Procede-se ao calculo do IP para avaliar a qualidade da IM.

```
#valores obs e imput de escol versus o IP

fit.escol <- with(imp,
glm(escol.na~idade+estcivilr+refeicoescat+snackr+origem+sx,family=bino
mial))

ps <- rowMeans(sapply(fit.escol$analyses, fitted.values))

escol.1 <- complete(imp,1)$escol

escol.na <- is.na(imp$data$escol)

xyplot(escol.1 ~ ps, groups=escol.na,xlab="Probabilidade de dados
omissos em escolaridade",

       ylab="Escolaridade",

         par.settings = simpleTheme(col.points = rep(c("blue","red")),

cex=1.2, pch=19),

         auto.key = list(columns=2, text = c("Observado","Imputado")))

#residuos de escol observado e escol imputado

fit <- lm(escol.1 ~ ps)

densityplot(~ residuals(fit), group = escol.na, plot.points = FALSE,

              ref = TRUE, xlab = "Resíduos da regressão de
escolaridade em função do IP",

            ylab="Densidade",

              scales = list(y=list(draw=F)),

              par.settings = simpleTheme(col.line =
rep(c("blue","red"))),

auto.key = list(columns = 2, text=c("Observado","Imputado")))

library(lattice)

com <- complete(imp, "long", inc=T)

col <- rep(c("blue","red")[1+as.numeric(is.na(imp$data$escol))],6)

stripplot(escol~.imp, data=com, jit=TRUE, fac=0.8, col=col,
pch=20,cex=1.4, xlab="Número de imputações",ylab="Escolaridade")
```

O modelo linear generalizado é calculado com base nas cinco bases de dados imputadas. A função pool compila os resultados, através das fórmulas de média e EP dos coeficientes, preconizadas por Rubin.

```
#modelo linear generalizado com bases de dados imputadas

fit <- with(imp,
glm(imc~escol+idade+estcivilr+anos+snackr,family=Gamma(link=identity))
)

a<-pool(fit)

a

summary(a)
```

9.1.9. Imputação múltipla por regressão linear não Bayesiana (Cenário 1)

Para proceder à IM, por RLN, basta alterar o método anterior (o PMM é o método *default*).

```
set.seed(145)

imp<-mice(dados,meth="norm.nob",print=F)

imp$imp$escol

head(imp$imp$escol)
```

9.1.10. Simulação para 20% dados omissos na escolaridade

Os 20% de dados omissos em escolaridade são calculadas a partir dos 6.3% já existentes, sendo apenas uma parte do total, retirada aleatoriamente.

```
#seleção aleatória de 14% para completar 20% de missing

linha<-1:nrow(dados)

linha<-linha[!is.na(dados$escol)]

set.seed(515)

amostra<-sample(linha,nrow(dados)*.14)

dados20<-dados

dados20$escol[amostra]<-NA

summary(dados20)

hist(dados20$escol,main="Histograma escolaridade (20% dados
omissos)",ylab="Frequência",xlab="Escolaridade")

hist(dados$escol,main="Histograma escolaridade (6.8% dados
omissos)",ylab="Frequência",xlab="Escolaridade")
```

```
retira<-complete.cases(dados20[,-4])#ficar sem dados omissos, exceto
os da coluna 4 (escolaridade)

dados20<-dados20[retira,]

dadoscompletos20<-dados20[complete.cases(dados20),] #dados completos,
sem omissos

summary(dados20)

dim(dadoscompletos20)

dim(dados20)
```

9.1.11. Simulação para 20% de dados omissos na idade

Para calcular 20% de dados omissos em idade, são retirados 20% dos dados observados de forma aleatória, podendo assim garantir-se que o processo de omissão é MCAR.

```
#MISSINGS 20% IDADE

#simulação para 20% missing na idade

linha<-1:nrow(dadoscompletos)

linha<-linha[!is.na(dadoscompletos$idade)]

set.seed(515)

amostra<-sample(linha,nrow(dados)*.20)

dados20<-dados

dados20$idade[amostra]<-NA

summary(dados20)

dim(dados20)

par( mfrow = c( 1, 1) )

hist(dados20$idade,main="Histograma idade após
simulação",xlab="Idade",ylab="Frequência")

hist(dados$idade,main="Histograma idade pré-
simulação",xlab="Idade",ylab="Frequência")

length(dados$idade)

summary(is.na(dados20$idade))

#dados completos

retira<-complete.cases(dados20[,-2])#ficar sem dados omissos, exceto
os da coluna 4 (escolaridade)

dados20<-dados20[retira,]
```

```
dadoscompletos20<-dados20[complete.cases(dados20),] #dados completos,
sem omissos

summary(dados20)

summary(dadoscompletos20)

dim(dados20)
```

9.2. Cópia das questões do inquérito usadas no presente estudo

Avaliação do Acesso aos Cuidados de Saúde e Nível de Saúde das Comunidades Imigrantes Africana e Brasileira em Portugal

Está a ser levado a cabo pelo Instituto de Medicina Preventiva, da Faculdade de Medicina de Lisboa, um estudo com o objectivo de caracterizar o estado de saúde e a prestação efectiva de cuidados de saúde às comunidades imigrantes a residir em Portugal.
Vimos deste modo solicitar a sua disponibilidade e colaboração para nos responder a algumas perguntas relativas ao seu estado de saúde, práticas e hábitos individuais, a fim de podermos alcançar um melhor conhecimento sobre esta realidade.
Espera-se com este projecto poder contribuir para o desenvolvimento de políticas de saúde e estratégias direccionadas para os imigrantes, no sentido de reduzir as desigualdades de saúde, no contexto do Plano Nacional de Saúde.
Todas as informações registadas neste questionário são ESTRITAMENTE CONFIDENCIAIS e apenas serão usadas de acordo com as finalidades deste inquérito. Por favor, seja sincero. Agradecemos, desde já, a sua colaboração!

I – CARACTERIZAÇÃO SÓCIO-DEMOGRÁFICA	I – CARACTERIZAÇÃO SÓCIO-DEMOGRÁFICA
0. SEXO Masculino.....1 Feminino.....2	☐
1. QUAL É A [SUA] DATA DE NASCIMENTO [DO(A) SR(A)_____]? _____ (Se não sabe: registe a data aproximada) A [sua] idade [do(a) sr(a)_____] é _____ anos	☐☐ ☐☐ ☐☐ Dia Mês Ano ☐☐ Idade
2. QUAL É O [SEU] ESTADO CIVIL [DO(A) SR(A)_____]? Solteiro(a).....1 Casado(a) ou em união de facto.....2 Divorciado(a) ou Separado(a).....3 Viúvo(a).....4 Não sabe.....9	☐
3. QUAL O NÍVEL DE ENSINO MAIS ELEVADO QUE [O(A) SR(A)_____] FREQUENTA OU QUE FREQUENTOU? Nenhum.....1 → P.5 Ensino básico – 1º ciclo.....2 Ensino básico – 2º ciclo.....3 Ensino básico – 3º ciclo.....4 Ensino secundário.....5 Ensino pós-secundário.....6 Ensino superior – Bacharelato.....7 Ensino superior – Licenciatura.....8 Ensino superior – Mestrado.....9 Ensino superior – Doutoramento.....10 Não sabe.....99 → P.5	☐☐

3/41

9. DIGA-ME, POR FAVOR, O N.º TOTAL DE HORAS SEMANAIS QUE [O(A) SR(A)_____] TRABALHA (TRABALHAVA). _____horas Não sabe.....99	□□ Horas
10. QUAL O TIPO DE ACTIVIDADE A QUE SE DEDICA (DEDICAVA) O ESTABELECIMENTO EM QUE [O(A) SR(A)_____] TRABALHA (TRABALHAVA)? _____ Não sabe.....99	□□

II – TRAJECTÓRIA MIGRATÓRIA	**II – TRAJECTÓRIA MIGRATÓRIA**
11. QUAL É A [SUA] NACIONALIDADE [DO(A) SR(A)_____]? Portuguesa.....1 Estrangeira (indique).....2 Dupla nacionalidade (indique).....3 Apátrida (sem nacionalidade).....4 Não sabe.....9 _____	□ □□ Código País
12. [O(A) SR(A)_____] É NATURAL DE QUE PAÍS/REGIÃO (EM QUE PAÍS/REGIÃO NASCEU)? Portugal.....1 Outro (indique o país E região).....2 Não sabe.....9 _____ (Se nasceu em Portugal → P.16)	□ □□ Código País □□-□□ Código Região
13. HÁ QUANTOS ANOS [O(A) SR(A)_____] RESIDE EM PORTUGAL? _____anos Menos de um ano.....000 Não sabe.....999	□□□ Anos
14. QUAL O PRINCIPAL MOTIVO QUE [O(A)] FEZ [O(A) SR(A)_____] VIR PARA PORTUGAL? Motivo económico.....1 → P.16 Perseguição política.....2 → P.16 Motivo profissional.....3 → P.16 Afinidades culturais.....4 → P.16 Motivo de estudos.....5 → P.16 Reunificação familiar.....6 → P.16 Motivo de saúde (indique qual o problema de saúde)...7 Outro motivo (indique).....8 → P.16 _____	□

15. [O(A) SR(A)_____] TEVE VISTO TEMPORÁRIO DE ENTRADA EM PORTUGAL PARA TRATAMENTO MÉDICO (JUNTA MÉDICA)? Sim.....1 Não.....2 Não responde.....8 Não sabe.....9	☐
16. QUAL É A [SUA] SITUAÇÃO ACTUAL EM TERMOS DE ESTADA EM PORTUGAL [DO(A) SR(A)_____]? Tem Bilhete de Identidade português.....1 Tem uma autorização de residência permanente.....2 Tem uma autorização de residência temporária.....3 Tem uma autorização de permanência.....4 Tem um visto de trabalho.....5 Tem um visto de estudo.....6 Tem um visto de estada temporária.....7 Está à espera de documentação.....8 Não tem a situação regularizada.....9 Outra (indique).....10 Não responde.....88 Não sabe.....99 _____	☐☐
17. QUAL É A NATURALIDADE DO [SEU] PAI [DO(A) SR(A)_____]? Portugal.....1 Outro (indique o país).....2 Não sabe.....9 _____	☐ ☐☐ Código País
18. QUAL É A NATURALIDADE DA [SUA] MÃE [DO(A) SR(A)_____]? Portugal.....1 Outro (indique o país).....2 Não sabe.....9 _____	☐ ☐☐ Código País
19. QUAL É A NATURALIDADE DO [SEU] AVÔ PATERNO [DO(A) SR(A)_____]? Portugal.....1 Outro (indique o país).....2 Não sabe.....9 _____	☐ ☐☐ Código País

114. COM QUE IDADE [O(A) SR(A)_____] DEIXOU DE FUMAR? _____anos (Se não sabe peça a idade aproximada)	☐☐ Anos

VIII – CONSUMO DE ALIMENTOS E BEBIDAS	VIII – CONSUMO DE ALIMENTOS E BEBIDAS
VOU FAZER ALGUMAS PERGUNTAS SOBRE O QUE AS PESSOAS COSTUMAM COMER E BEBER. POR REFEIÇÕES PRINCIPAIS ENTENDE-SE O PEQUENO-ALMOÇO, O ALMOÇO E O JANTAR. **115. QUANTAS REFEIÇÕES PRINCIPAIS É QUE [O(A) SR(A)_____] TOMA HABITUALMENTE POR DIA?** _____refeições Não sabe……9	☐ Refeições
116. O QUE [O(A) SR(A)_____] COMEU ONTEM NAS 3 REFEIÇÕES PRINCIPAIS? A. Leite/iogurte/queijo……..Sim Não NS B. Sopa……..Sim Não NS C. Pão……..Sim Não NS D. Carne……..Sim Não NS E. Peixe……..Sim Não NS F. Batatas/arroz/massa……..Sim Não NS G. Feijão/grão……..Sim Não NS H. Salada/legumes cozidos……..Sim Não NS I. Fruta……..Sim Não NS J. Bolos/chocolates/sobremesa doce……..Sim Não NS K. Outros alimentos (Indique quais)……..Sim Não NS _____ (Se "Não" em todas as alternativas, assinale "Sim" em "Não comeu nada"; caso contrário, assinale "Não") L. Não comeu nada……..Sim Não NS	A.☐ B.☐ C.☐ D.☐ E.☐ F.☐ G.☐ H.☐ I.☐ J.☐ K.☐ L.☐ Sim…….. 1 Não…….. 2 Não sabe… 9
117. [O(A) SR(A)_____] COME HABITUALMENTE FORA DAS 3 REFEIÇÕES PRINCIPAIS? Se sim: **QUANTAS VEZES POR DIA** _____vezes Não come fora das refeições……..00 Não sabe……..99	☐☐ Vezes

9.3. Relação linear entre variável resposta e variáveis contínuas explicativas (modelo linear generalizado dos casos completos, cenário 1)

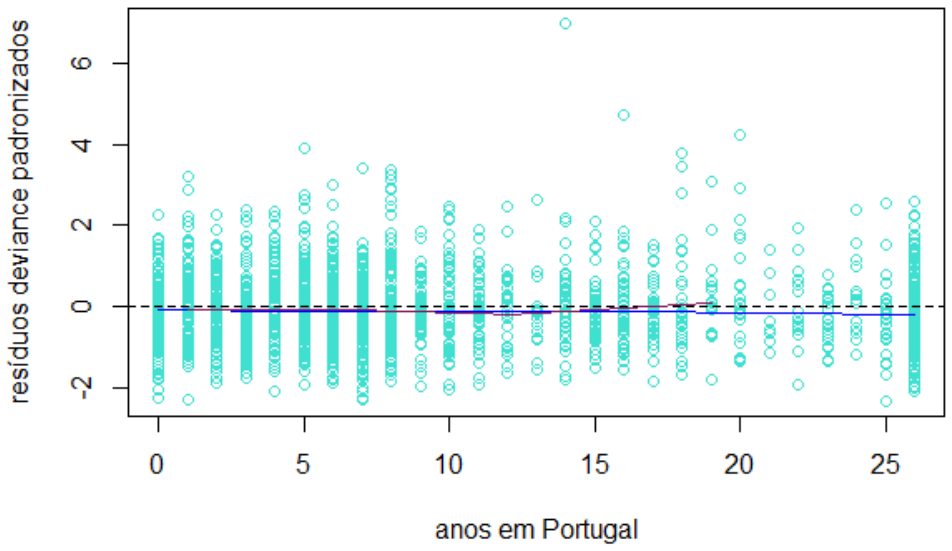


Figura 53 - Relação linear entre IMC e anos (resíduos versus anos)

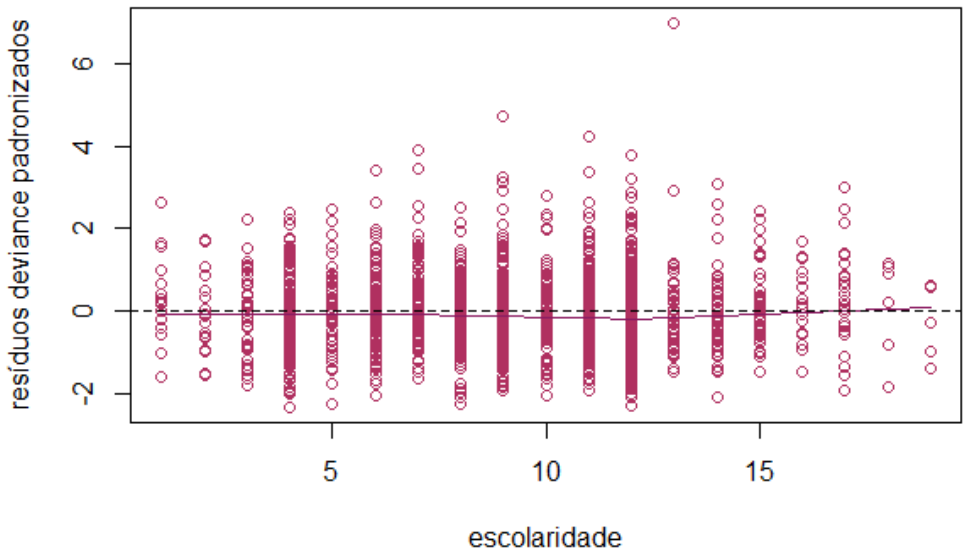


Figura 54 - Relação linear entre IMC e escolaridade (resíduos versus escolaridade)

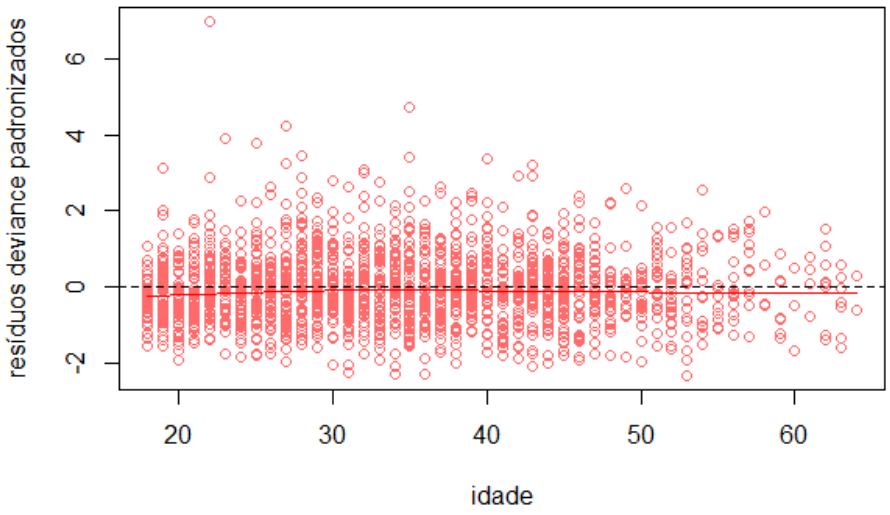


Figura 55 - Relação linear entre IMC e idade (resíduos versus idade)

9.4. Gráficos da distribuição marginal

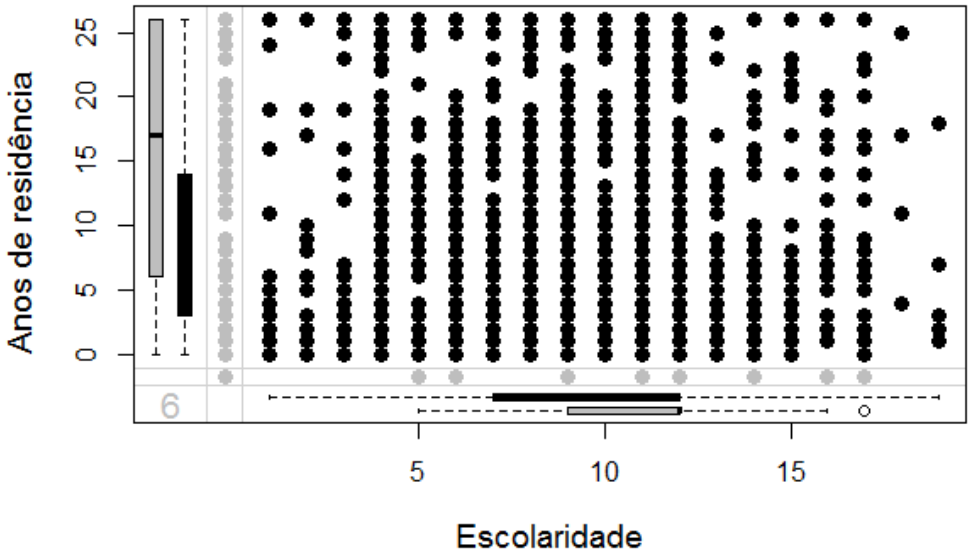


Figura 56 - Gráfico da distribuição marginal dos anos *versus* escolaridade. Dados observados a preto e dados omissos a cinzento

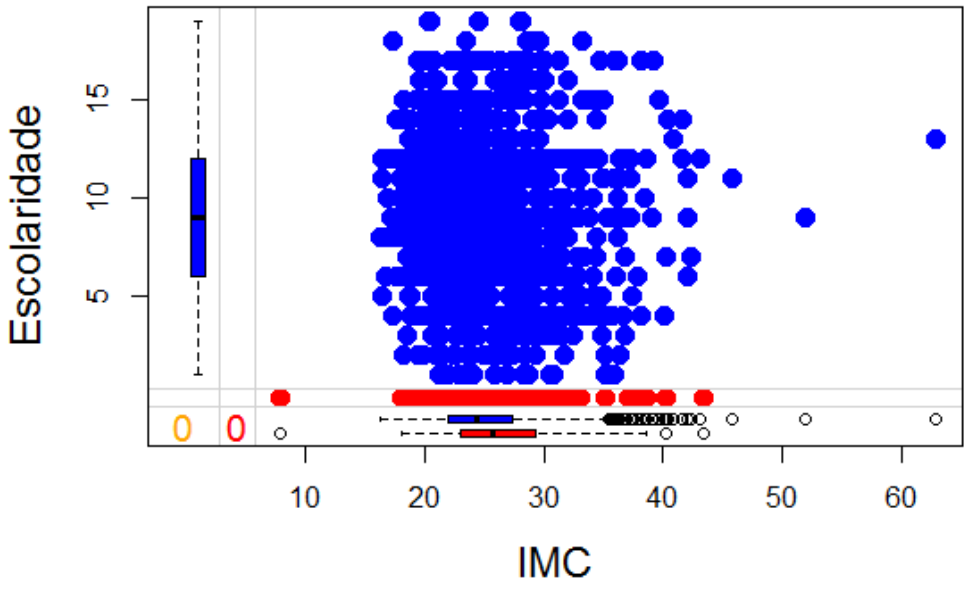


Figura 57 - Gráfico da distribuição marginal dos IMC *versus* escolaridade. Dados observados a preto e dados omissos a cinzento

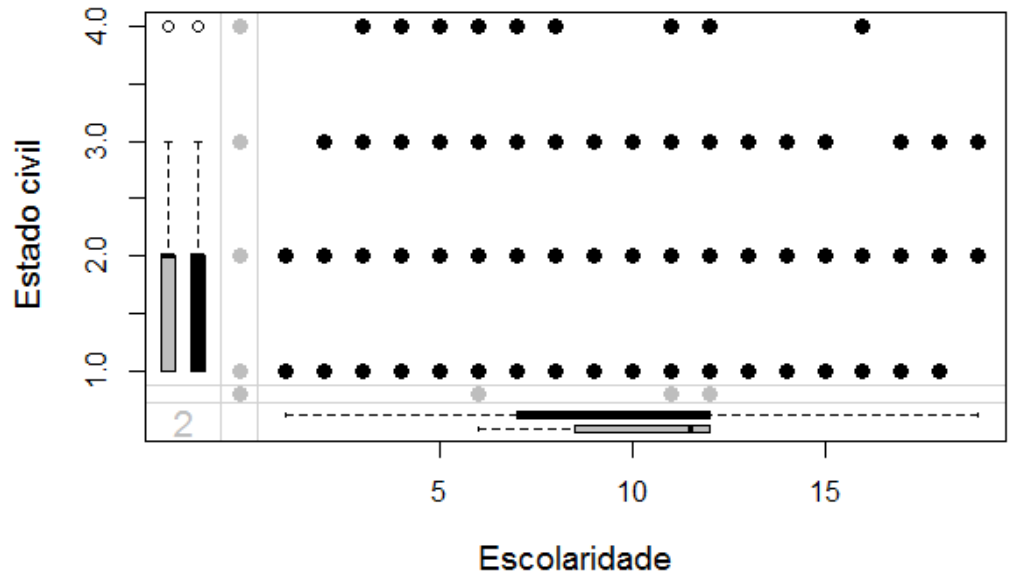


Figura 58 - Gráfico da distribuição marginal dos estado civil *versus* escolaridade. Dados observados a preto e dados omissos a cinzento

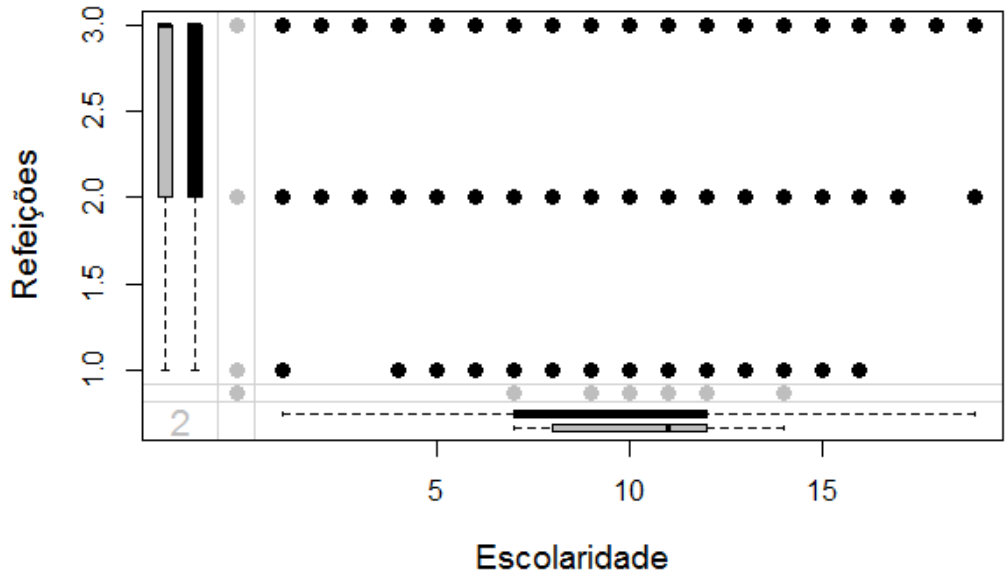


Figura 59 - Gráfico da distribuição marginal dos refeições *versus* escolaridade. Dados observados a preto e dados omissos a cinzento

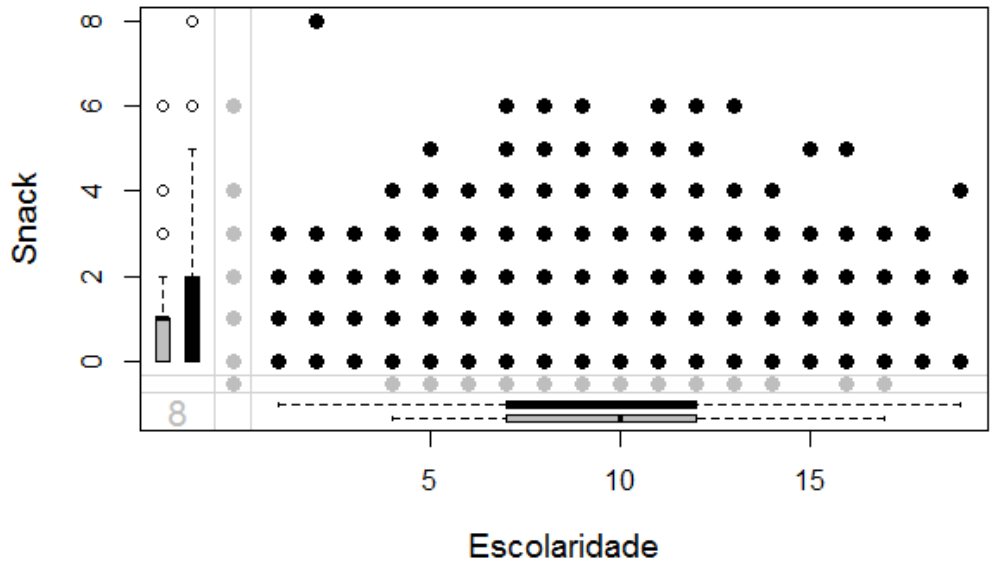


Figura 60 - Gráfico da distribuição marginal dos *snack versus* escolaridade. Dados observados a preto e dados omissos a cinzento

9.5. Sumário das variáveis imputadas em cada base de dados imputada

1	2	3	4	5
Min. : 1.000	Min. : 2.000	Min. : 1.000	Min. : 1.000	Min. : 1.000
1st Qu.: 4.000	1st Qu.: 4.000	1st Qu.: 4.000	1st Qu.: 4.000	1st Qu.: 4.000
Median : 6.000	Median : 6.000	Median : 6.000	Median : 5.000	Median : 6.000
Mean : 6.924	Mean : 7.202	Mean : 7.286	Mean : 6.798	Mean : 6.756
3rd Qu.: 9.000	3rd Qu.:11.000	3rd Qu.:11.500	3rd Qu.: 9.500	3rd Qu.:11.000
Max. :15.000	Max. :18.000	Max. :17.000	Max. :18.000	Max. :19.000

Figura 61 - Sumário da variável escolaridade após IM por PMM em cada base dados imputada (1 a 5) (Cenário 1)

1.v1	2.v1	3.v1	4.v1
Min. :-4.47932002396	Min. :-1.29880497125	Min. :-1.18040373365	Min. :-3.98480111721
1st Qu.: 5.11633679254	1st Qu.: 5.01537903816	1st Qu.: 3.33732888218	1st Qu.: 3.52806378269
Median : 7.10162312221	Median : 7.41229194889	Median : 6.54029422044	Median : 6.83735332297
Mean : 7.44153020010	Mean : 7.42487266012	Mean : 6.48070498619	Mean : 6.64695329997
3rd Qu.: 9.79035581730	3rd Qu.:10.16133312120	3rd Qu.: 9.08465243007	3rd Qu.: 9.61556269937
Max. :16.08378236120	Max. :19.30420998190	Max. :16.51487347610	Max. :14.34834164660

5.v1
Min. : 0.121145488047
1st Qu.: 5.481409618290
Median : 8.074674152750
Mean : 7.975276988080
3rd Qu.:10.565112067900
Max. :16.846430389200

Figura 62 - Sumário da variável escolaridade após IM por RLN em cada base dados imputada (1 a 5) (Cenário 1)

1	2	3	4	5
Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 1.000
1st Qu.: 5.000	1st Qu.: 5.000	1st Qu.: 5.000	1st Qu.: 4.000	1st Qu.: 5.750
Median : 9.000	Median : 9.000	Median : 9.000	Median : 9.000	Median : 9.000
Mean : 8.646	Mean : 8.576	Mean : 8.534	Mean : 8.456	Mean : 8.836
3rd Qu.:12.000	3rd Qu.:12.000	3rd Qu.:12.000	3rd Qu.:12.000	3rd Qu.:12.000
Max. :19.000	Max. :19.000	Max. :19.000	Max. :18.000	Max. :19.000

Figura 63 - Sumário da variável escolaridade após IM por PMM em cada base dados imputada (1 a 5) (Cenário 2)

1	2	3	4	5
Min. :18.00	Min. :18.00	Min. :18.00	Min. :18.00	Min. :18.00
1st Qu.:26.00	1st Qu.:26.00	1st Qu.:26.00	1st Qu.:26.00	1st Qu.:26.00
Median :34.00	Median :34.00	Median :34.00	Median :34.00	Median :34.00
Mean :34.75	Mean :34.54	Mean :34.82	Mean :34.77	Mean :34.76
3rd Qu.:43.00	3rd Qu.:41.00	3rd Qu.:42.00	3rd Qu.:42.00	3rd Qu.:43.00
Max. :62.00	Max. :63.00	Max. :63.00	Max. :62.00	Max. :63.00

Figura 64 - Sumário da variável escolaridade após IM por PMM em cada base dados imputada (1 a 5) (Cenário 3)

9.6. Stripplots das variáveis após IM

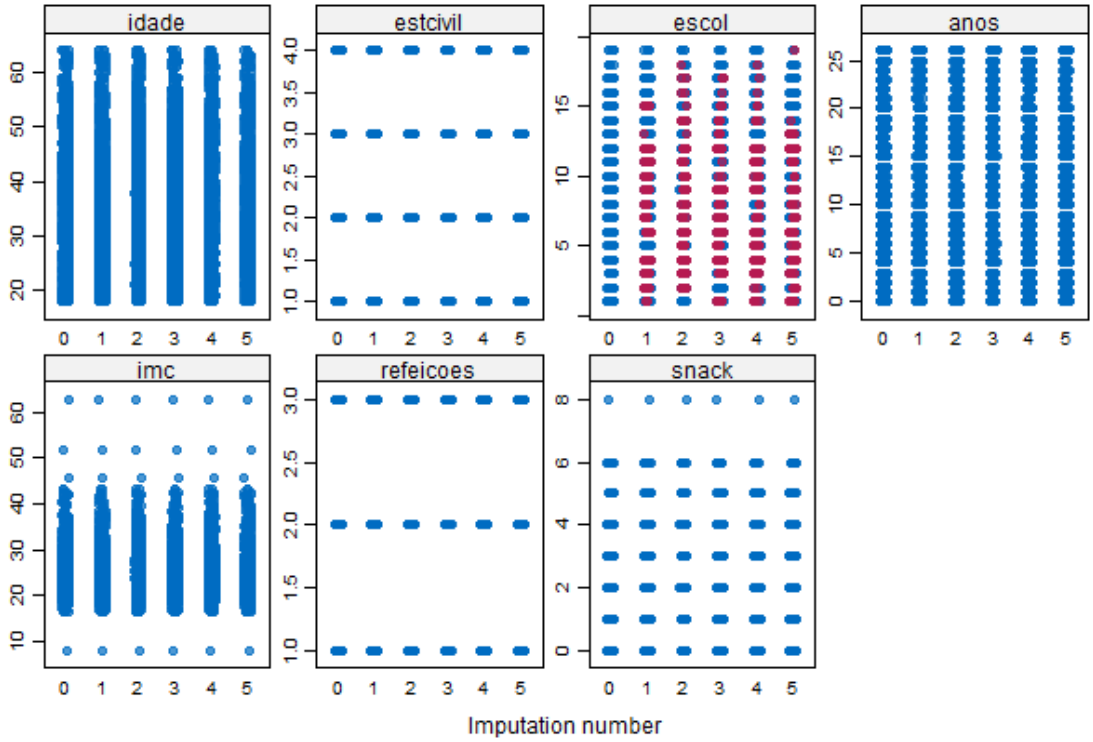


Figura 65 - Stripplot das variáveis antes e depois da IM (a azul valores observados e a vermelho valores imputados) - IMM por PMM Cenário 1

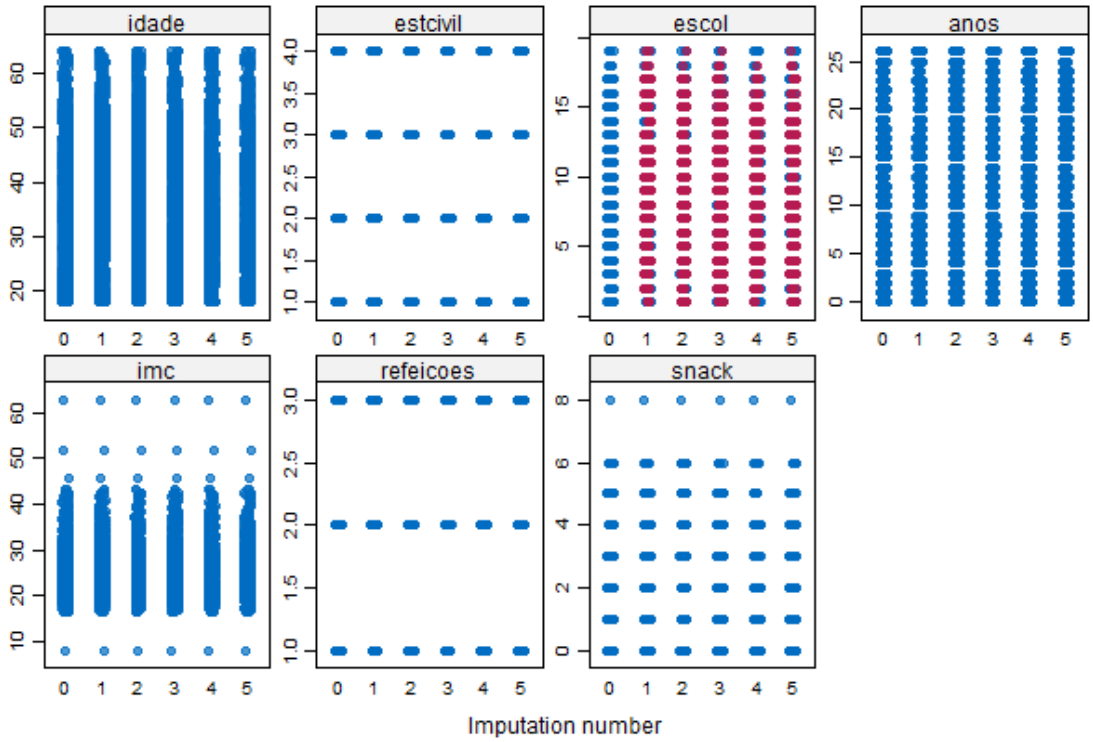


Figura 66 - Stripplot de todas as variáveis antes e depois da IM (a azul valores observados e a vermelho valores imputados) - Cenário 2

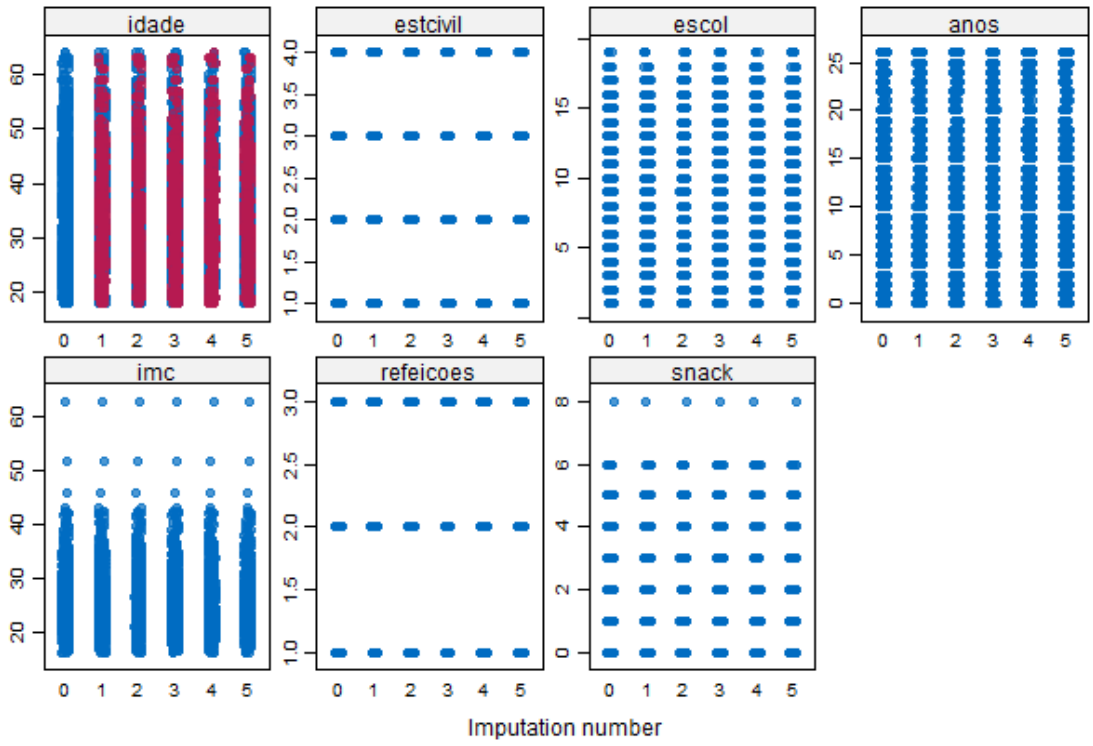


Figura 67 - Stripplot de todas as variáveis antes e depois da IM (a azul valores observados e a vermelho valores imputados) - Cenário 3

9.7. Representação gráfica da convergência das iterações na IM

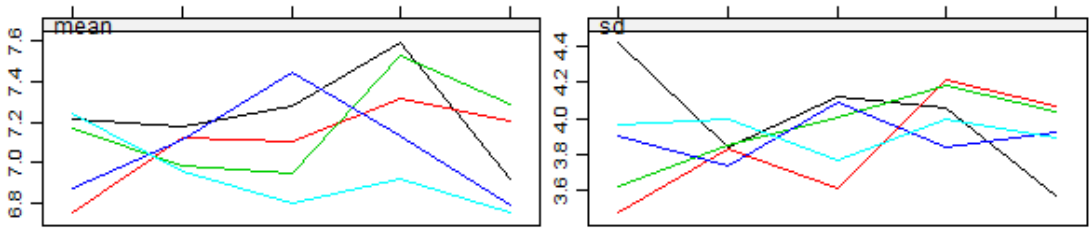


Figura 68 - Linhas de convergência das iterações IM por PMM (Cenário 1)

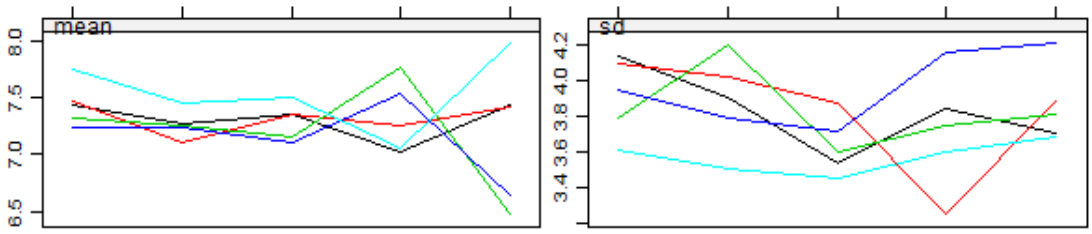


Figura 69 - Linhas de convergência das iterações IM por RLN (Cenário 1)

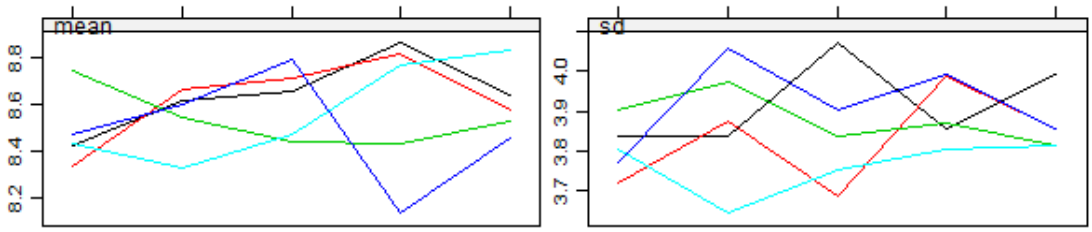


Figura 70 - Linhas de convergência das iterações IM por PMM (Cenário 2)

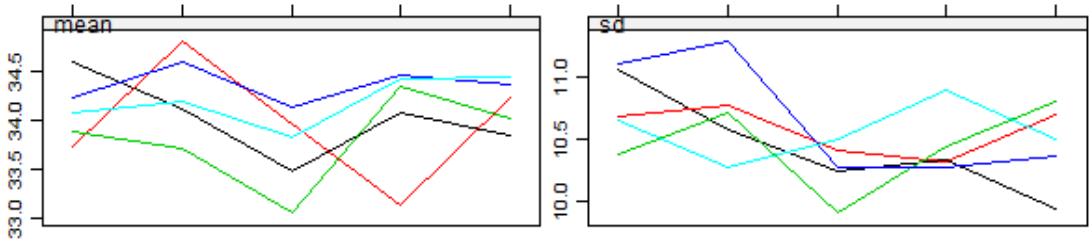


Figura 71 - Linhas de convergência das iterações IM por PMM (Cenário 3)