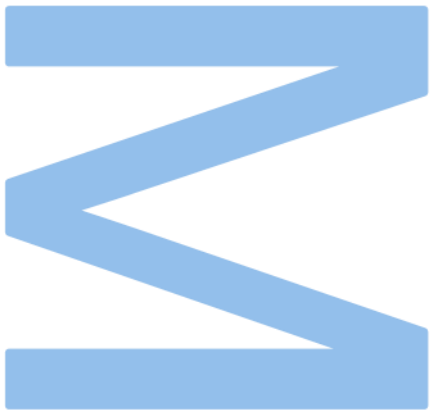
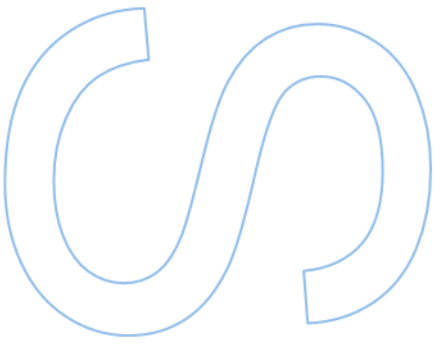
Imputação múltipla na modelação do atraso do diagnóstico na Tuberculose

Maria Francisca Moreira de Barros
Mestrado de Estatística Computacional e Análise de Dados
Departamento de Matemática
2023

Supervisor
Óscar Felgueiras, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO





Agradecimentos

Em primeiro lugar, gostaria de expressar a minha profunda gratidão ao meu orientador, Professor Óscar Felgueiras, pela sua orientação, paciência e apoio ao longo deste ano letivo.

Agradeço também ao Instituto de Saúde Pública da Universidade do Porto (ISPUP) pela oportunidade de realizar o meu estágio curricular, em particular à Professora Dra. Raquel Duarte.

Estou igualmente grata a todos os meus colegas e amigos que me apoiaram, encorajaram e acompanharam ao longo deste percurso académico.

Um agradecimento especial à minha família, pelo apoio incondicional, compreensão e paciência ao longo de todos estes anos de estudo.

Por último, mas não menos importante, gostaria de agradecer a todos aqueles que contribuíram, direta ou indiretamente, para a realização deste relatório de estágio.

Resumo

Este relatório centra-se na investigação do atraso no diagnóstico da tuberculose, em particular nos cuidados de saúde, em Portugal Continental entre 2008 e 2017. A tuberculose, uma infeção causada pela bactéria *Mycobacterium tuberculosis*, tem a sua transmissão mais prevalente entre o início dos sintomas e o início do tratamento, tornando o diagnóstico atempado essencial para prevenir a sua propagação.

O estudo aborda o impacto dos valores omissos em análises estatísticas, classificando-os e discutindo os métodos de imputação, com destaque para o método de imputação múltipla usando o pacote `mice` do software `R`. A metodologia central do relatório envolve a aplicação de Modelos Aditivos Generalizados (GAM), modelos estatísticos flexíveis que permitem capturar relações complexas entre as variáveis.

A aplicação prática deste estudo foi conduzida utilizando a base de dados do Sistema Nacional de Vigilância da Tuberculose em Portugal, o SVIG-TB. Inicialmente foi realizada uma análise exploratória dos dados, seguida pela modelagem estatística. Esta modelagem revelou várias associações significativas entre as variáveis em estudo, como o sexo, a idade e outros fatores de risco, e o tempo de atraso no diagnóstico.

O relatório destaca a importância de uma abordagem cuidadosa ao tratar de dados omissos em análises estatísticas, sobretudo em contextos clínicos e de saúde pública. As descobertas deste estudo são cruciais para entender os fatores associados ao atraso no diagnóstico da tuberculose e para a formulação de estratégias de intervenção eficazes na área da saúde pública.

Palavras-chave: Tuberculose, imputação múltipla, R, modelos aditivos generalizados, tempo de atraso do diagnóstico nos cuidados de saúde.

Abstract

This report focuses on the investigation of delays in the diagnosis of tuberculosis, particularly in healthcare, in mainland Portugal between 2008 and 2017. Tuberculosis, an infection caused by the bacterium *Mycobacterium tuberculosis*, has its most prevalent transmission between the onset of symptoms and the start of treatment, making timely diagnosis essential to prevent its spread.

The study addresses the impact of missing values in statistical analyses, classifying them and discussing imputation methods, with an emphasis on the multiple imputation method using the `mice` package in the `R` software. The core methodology of the report involves the application of Generalized Additive Models (GAM), flexible statistical models that allow capturing complex relationships between variables.

The practical application of this study was conducted using the database of the National Tuberculosis Surveillance System in Portugal, the SVIG-TB. An initial exploratory analysis of the data was followed by statistical modeling. This modeling revealed several significant associations between the study variables, such as gender, age, and other risk factors, and the delay in diagnosis.

The report highlights the importance of a careful approach when dealing with missing data in statistical analyses, especially in clinical and public health contexts. The findings of this study are crucial for understanding the factors associated with delays in the diagnosis of tuberculosis and for formulating effective intervention strategies in the field of public health.

Keywords: Tuberculosis, multiple imputation, R, generalized additive models, delay in diagnosis in healthcare.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O presente relatório resultou de um estágio curricular desenvolvido no Instituto de Saúde Pública da Universidade do Porto (ISPUP). Durante este estágio, sob a orientação do Professor Óscar Felgueiras e da Professora Dra. Raquel Duarte, tive o privilégio de trabalhar num projeto sobre o atraso no diagnóstico da tuberculose.

A tuberculose (TB) é uma doença infeciosa causada pela *Mycobacterium tuberculosis*, cuja forma de apresentação mais frequente é a tuberculose pulmonar. Os indivíduos infetados podem transmitir a doença, o que resulta na propagação da bactéria e consequente formação e continuidade de uma cadeia de transmissão [1]. O diagnóstico precoce é um fator chave para impedir a propagação desta doença, uma vez que a janela de contágio situa-se entre o início dos sintomas e o início do tratamento [2].

O atraso da diagnosticação da tuberculose corresponde ao período desde o início dos sintomas até à data do diagnóstico, podendo ser dividido em dois períodos: o atraso relacionado com o doente (que corresponde ao tempo desde o início dos sintomas até à primeira vez que o doente procura ajuda médica) e o atraso relacionado com o serviço de saúde (período com início no primeiro contacto entre o profissional de saúde e o doente, e término no momento do diagnóstico da doença) [3].

A maioria dos estudos realizados neste tema foram feitos em países de elevada incidência da tuberculose, com recursos limitados [4] [5]. No entanto, existem registos de atrasos consideráveis no diagnóstico em países com baixa incidência e mais recursos [6] [7] [8] [9].

Existe um leque variado de determinantes do atraso no diagnóstico da TB, tais como o género e idade (por norma, as mulheres e os idosos têm atrasos maiores), determinados fatores sociais (desemprego, baixo salário e baixo nível de escolaridade sobretudo em

contextos de recursos reduzidos-médios) e a nacionalidade (estudos realizados em países com mais recursos mostraram que os doentes naturais do país tinham maior atraso no diagnóstico) [6] [7] [8] [9] [10].

O objetivo do nosso trabalho foi avaliar o atraso médio no diagnóstico entre o período de 2008 a 2017 em Portugal Continental, mais especificamente o atraso no diagnóstico associado aos cuidados de saúde, e compreender quais os seus principais determinantes.

1.1 Métodos utilizados

Neste estudo, optamos por utilizar Modelos Aditivos Generalizados (GAM) para analisar o atraso no diagnóstico da tuberculose nos cuidados de saúde. Os GAM são ferramentas poderosas e flexíveis na modelagem estatística que combinam as propriedades dos Modelos Lineares Generalizados (GLM) e dos Modelos Aditivos (AM) [11]. Este modelo foi implementado utilizando o software `R`, através do pacote `mgcv`.

Os GAM são particularmente adequados no contexto do nosso estudo devido à relação entre as variáveis explicativas e a resposta ser não linear. Deste modo, o uso de Modelos Lineares Generalizados torna-se inadequado, uma vez que os pressupostos não seriam satisfeitos.

A base de dados utilizada neste estudo apresenta valores omissos em algumas variáveis. Os valores omissos em conjuntos de dados podem levar a inferências estatísticas enviesadas se não forem tratados corretamente. Assim, utilizamos o método de imputação múltipla, implementado através do pacote `mice` do software `R`. Este método substitui cada valor ausente por um conjunto de valores plausíveis, criando múltiplas versões completas do conjunto de dados para análise [12].

A imputação múltipla permite uma análise mais robusta, levando em consideração a incerteza associada à imputação dos dados omissos. Deste modo, conseguimos maximizar o uso de todas as informações disponíveis no nosso conjunto de dados, melhorando a precisão e a confiabilidade das nossas conclusões.

1.2 Estrutura do Relatório de Estágio

O relatório está organizado em cinco capítulos, no Capítulo 1 é introduzido o contexto da tuberculose, a relevância de um diagnóstico atempado e as consequências de atrasos

neste processo, além de delinear os objetivos principais do estudo e a metodologia adotada.

No Capítulo 2 é introduzido o conceito de valores omissos, abordando o impacto e as razões para a omissão de dados em análises estatísticas. Exploramos os métodos de imputação para lidar com dados omissos, destacando o método de imputação múltipla, em particular o mice, que é fundamental para a compreensão e tratamento dos dados utilizados no estudo.

O Capítulo 3 é dedicado à descrição detalhada dos Modelos Aditivos Generalizados (GAM), este capítulo fornece uma compreensão dos GAM, abordando a sua estrutura, a representação e estimação de funções suaves e a relevância da suavidade na modelagem estatística.

Os resultados obtidos ao longo do estudo são apresentados no Capítulo 4. O capítulo faz uma descrição detalhada da organização da base de dados, seguida de uma análise exploratória e, por fim, a aplicação prática dos métodos abordados nos capítulos anteriores isto é, do processo de imputação múltipla e da aplicação de modelos e interpretação dos mesmos.

Por fim, o Capítulo 5 resume as principais conclusões do estudo e discute as limitações encontradas, bem como futuros tópicos a ser abordados.

Capítulo 2

Valores Omissos

2.1 Introdução

O risco de enviesamento devido a valores omissos depende das razões pelas quais os dados estão em falta. Estas razões para valores omissos podem ser classificadas como:

1. ***Missing completely at random*** (MCAR): A ausência de dados não tem nenhuma relação com os mesmos, ou seja, a razão para a falta de dados é independente do que é observado e do que não é observado. Um exemplo disso ocorre quando escolhemos uma amostra aleatória de uma população, onde cada indivíduo tem a mesma probabilidade de ser incluído na amostra. Os dados (não observados) dos indivíduos da população que não foram incluídos na amostra são MCAR. Em particular, seguindo o exemplo dado por Kleinke et al. [13], vamos supor que temos um conjunto de dados com duas variáveis em que existem valores omissos numa delas, *wage*, e a outra, *nchild*, é totalmente observada. Os valores em falta de *wage* são MCAR se a probabilidade de observar ou não observar um valor de *wage* é a mesma para todos os possíveis valores da variável *nchild* e para todos os valores observados de *wage*.

2. ***Missing at random*** (MAR): A probabilidade de um dado estar ausente pode ser explicada por outras variáveis observadas no conjunto de dados. Assim, quando escolhemos uma amostra de uma população, os valores omissos são MAR quando a probabilidade de ser incluído depende de alguma propriedade conhecida. MAR é uma suposição mais flexível e realista do que MCAR. Em particular, seguindo o exemplo anterior, vamos

agora assumir que as duas variáveis estão negativamente correlacionadas, ou seja se indivíduos com um menor número de filhos tendem a não responder à pergunta sobre o salário. Neste caso, os valores omissos de *wage* são MAR, uma vez que dentro de cada classe definida pelo número de filhos, a probabilidade de um valor salarial estar em falta não altera com os possíveis valores de *wage*. Assim, podemos dizer que os valores omissos de *wage* são MCAR condicionados ao número de filhos que vivem na mesma casa.

3. ***Missing not at random*** (MNAR): Mesmo após levar em conta os dados observados, ainda existem diferenças sistemáticas entre os valores em falta e os valores observados, ou seja, o valor omisso está relacionado com a razão pela qual está em falta. Um exemplo de MNAR no contexto de uma pesquisa de opinião pública ocorre quando aqueles com opiniões mais fracas respondem com menos frequência. MNAR é o caso mais complexo. Considerando novamente as duas variáveis *wage* e *nchild*. Se indivíduos com salários altos tendem a não responder à pergunta salarial, mesmo dentro do grupo com o mesmo número de filhos, então os valores omissos seriam do tipo MNAR.

2.2 Métodos de Imputação

Os métodos de imputação são métodos que geram uma ou mais predições para cada valor em falta. Os métodos de imputação podem ser adotados se os dados omissos forem MCAR, MAR ou MNAR. No entanto, no último caso são necessárias suposições fortes ou informações externas devem estar disponíveis para permitir inferências válidas.

Os pacotes de software que fornecem ferramentas de imputação geralmente pressupõem que os valores omissos são MAR.

Assim, o mais importante é perceber como gerar imputações de forma a que inferências de interesse científico tendam a ser válidas mesmo na presença de dados omissos.

2.2.1 Imputação Simples

Neste método, substitui-se os dados omissos de uma variável Y pela sua melhor previsão. Essa previsão depende não só do conjunto de dados, mas também da distribuição de Y. Por exemplo, se não houver covariáveis, a melhor previsão de Y pode ser a média ou a mediana observada, dependendo da distribuição da variável. Por outro lado, se houver covariáveis, a melhor previsão pode ser um modelo de regressão, ou outro modelo

de previsão. Uma vez concluída a imputação, os valores imputados são tratados como valores observados, verdadeiros e conhecidos.

No caso de conjuntos de dados de pouca dimensão ou com pouca percentagem de valores omissos, este método produz bons resultados. No entanto, para conjuntos de dados com uma grande percentagem de valores omissos este método pode causar um grande viés, resultando numa grande diferença entre os valores estimados e os valores reais. Deste modo, nestes casos é recomendado o uso da imputação múltipla.

2.2.2 Imputação Múltipla

A imputação múltipla é um método proposto para casos em que há uma proporção de valores omissos entre 30% a 50%, segundo Rubin [12]. A ideia geral é gerar múltiplas previsões ou imputações para cada valor omisso, criando assim múltiplas versões de um conjunto de dados que podem ser analisadas usando métodos padrão.

A imputação múltipla é um método que fornece inferências estatísticas válidas sob a condição MAR [14]. Este método reconhece a incerteza associada aos valores imputados gerando um conjunto de m valores plausíveis para cada ponto de dados não observado, resultando em m conjuntos de dados completos, cada um com uma estimativa única dos valores omissos. Em seguida, os m conjuntos de dados completos são analisados individualmente usando procedimentos estatísticos, resultando em estimativas ligeiramente diferentes para cada parâmetro. Na etapa final da imputação múltipla, as m estimativas são combinadas para produzir uma única estimativa do parâmetro e do seu correspondente erro padrão.

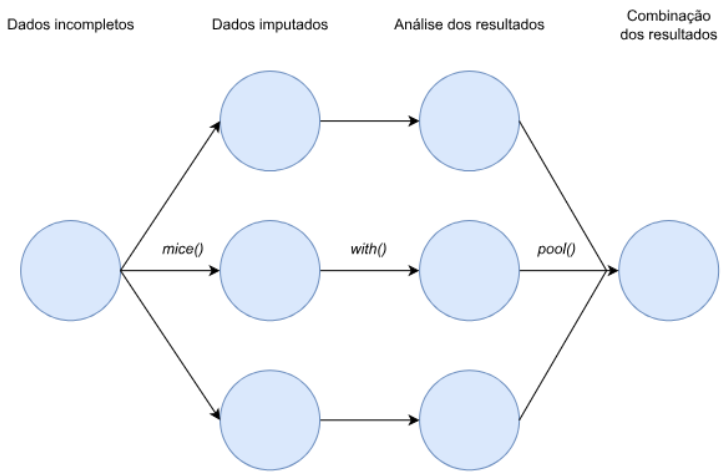


FIGURA 2.1: Principais etapas da imputação múltipla para $m = 3$. Figura adaptada de [15]

Em resumo, este método consiste em três etapas ilustradas na Figura 2.1. Primeiro, imputar os dados omissos m vezes para produzir m conjuntos de dados completos; de seguida, analisar cada conjunto de dados usando um procedimento estatístico padrão; e, por fim, combinar os m resultados num único resultado.

2.2.2.1 Imputação

A etapa da imputação é a etapa mais complicada entre as três. O objetivo desta etapa é preencher os valores em falta várias vezes, usando as informações contidas nos dados observados. Existem vários métodos de imputação disponíveis, neste caso iremos focar-nos na imputação multivariada por equações encadeadas (MICE).

Imputação Multivariada por Equações Encadeadas

MICE é um algoritmo iterativo que imputa valores omissos com base em modelos de regressão. O algoritmo passa pelos seguintes passos, adaptados de Azur et al. [16] :

1. Todos os valores omissos são temporariamente preenchidos por imputação simples;
2. Para uma variável X todos os valores imputados temporariamente são eliminados, e é realizada uma regressão entre X e as outras variáveis presentes no conjunto de dados, sendo que X é a variável dependente neste modelo;
3. Os valores omissos de X são preenchidos com as previsões resultantes do modelo de regressão;
4. Os passos 2-3 são repetidos para todas as variáveis com valores omissos. Quando todas as variáveis são imputadas é considerado o fim de um ciclo de imputação;
5. Os passos 2-4 são repetidos para um número definido de ciclos, com as imputações atualizadas a cada ciclo.

A vantagem deste método é ter a possibilidade de ajustar o modelo de regressão para cada variável, e assim ser capaz de estimar todos os diferentes tipos de variáveis, contínuas ou categóricas. No final do passo 5, as imputações finais são mantidas, resultando num conjunto de dados imputados. O algoritmo completo é então repetido m vezes, criando m conjuntos de dados imputados em que os valores originalmente omissos diferem.

2.2.2.2 Análise e Combinação de resultados

A segunda etapa da imputação múltipla analisa os m conjuntos de dados separadamente usando um procedimento estatístico. No final desta etapa, são obtidos m conjuntos de estimativas de parâmetros a partir de análises separadas dos m conjuntos de dados.

A última etapa combina as m estimativas numa única estimativa. Rubin [12] forneceu fórmulas para combinar as m estimativas pontuais e os erros padrões numa única estimativa pontual e no seu erro padrão.

A ideia principal da imputação múltipla é que valores plausíveis possam ser usados no lugar dos valores omissos de maneira a que permitam que as estimativas de parâmetros não sejam enviesadas e, que a incerteza da estimativa de parâmetros no caso de dados omissos seja estimada de maneira razoável. A estimativa de um parâmetro, por exemplo, o coeficiente de regressão b, é simplesmente a média da estimativa do parâmetro obtida nos m conjuntos de dados, dada por

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i \tag{2.1}$$

onde b_i é o coeficiente de regressão para o i-ésimo conjunto de dados. No entanto, a variância é dada pela soma ponderada de duas variâncias, a variância dentro de cada imputação U_b, que captura a variabilidade usual de amostragem, e a variância entre imputações B_b, que captura a variabilidade devido aos dados omissos. As seguintes fórmulas apresentadas foram retiradas de Graham et al. [17],

$$U_b = \frac{1}{m}\sum_{i=1}^{m} SE_{b_i}^2 \tag{2.2}$$

$$B_b = \frac{1}{m-1}\sum_{i=1}^{m} (b_i - \bar{b})^2 \tag{2.3}$$

onde $SE_{b_i}^2$ é o erro padrão ao quadrado para o coeficiente b do i-ésimo conjunto de dados. Então, combinando estas duas variâncias obtemos a variância combinada dada pela seguinte expressão,

$$T_b = U_b + \left(1 + \frac{1}{m}\right) B_b \tag{2.4}$$

e $SE_b = \sqrt{T_b}$. A estimativa do parâmetro é então dividida pelo seu SE para fornecer o valor t. Os graus de liberdade para este valor t são definidos por

$$df = (m-1)\left[1 + \frac{m \times U_b}{m+1} B_b\right]^2 \tag{2.5}$$

2.2.2.3 Medidas de informação de dados omissos

As medidas de informação de dados omissos ajudam a avaliar a eficácia da imputação múltipla ao lidar com os dados omissos, estas medidas incluem a fração de informação omitida (FMI), o aumento relativo na variância devido à falta de resposta (RIV) e a eficiência relativa (RE) [18]. Estas medidas são calculadas com base nos valores da variância dentro de cada imputação (U_b) e da variância entre imputações (B_b).

Em particular, a fração de informação omitida, FMI, fornece informações sobre a quantidade de informação ausente nos dados e o impacto da não resposta na variância dos resultados. Um valor de FMI perto de zero sugere que os dados omissos têm um impacto mínimo na estimativa de parâmetros, enquanto um valor de FMI próximo de 1 sugere que a informação omitida é substancial. Esta medida é calculada através da seguinte fórmula,

$$FMI = \frac{RIV + \frac{2}{df+3}}{RIV + 1} \tag{2.6}$$

onde RIV é o aumento relativo na variância devido aos dados ausentes, dado por,

$$RIV = \frac{(1 + m^{-1})B_b}{U_b} \tag{2.7}$$

Este parâmetro, FMI, é particularmente importante para medir a eficiência relativa de uma estimativa. A eficiência de uma estimativa, segundo Rubin [12], baseada em m imputações é, aproximadamente, dada por,

$$RE = \left(1 + \frac{FMI}{m}\right)^{-1} \tag{2.8}$$

A eficiência relativa fornece informações sobre a precisão da estimativa do parâmetro. Por exemplo, se um conjunto de dados tiver 50% de informação omitida, com $m = 5$

imputações, a eficiência da estimativa é 90.9%. Se aumentarmos o número de imputações geradas para $m = 10$, iremos obter uma eficiência de 95.2%.

Recomenda-se que o número de imputações seja suficientemente grande para garantir que as estimativas sejam estáveis e precisas. Normalmente, um número mínimo de 5 a 10 imputações é considerado suficiente para a maioria das situações. No entanto, em casos de dados com muitos valores omissos, pode ser necessário aumentar o número de imputações.

Uma maneira de avaliar se o número de imputações é suficiente é realizar uma análise de sensibilidade, na qual o número de imputações é aumentado gradualmente e as estimativas são comparadas para verificar se há mudanças significativas. Se não houver mudanças significativas nas estimativas, o número atual de imputações pode ser considerado suficiente.

Além disso, é importante lembrar que o aumento do número de imputações também se reflete no aumento do tempo e do custo computacional.

Capítulo 3

Modelo Aditivo Generalizado

3.1 Introdução

Os Modelos Aditivos Generalizados (GAM) foram originalmente desenvolvidos por Hastie and Tibshirani [11], com o objetivo de combinar as propriedades dos Modelos Lineares Generalizados (GLM) com os Modelos Aditivos.

Os GAM são modelos estatísticos não paramétricos que se concentram em modelar relações complexas entre variáveis dependentes e independentes, combinando funções lineares e não lineares para representar essas relações.

Ao contrário dos modelos lineares, os GAM permitem a adição de funções complexas, como *splines*, para capturar relações não lineares de maneira mais eficiente, sem a necessidade de se ajustarem a modelos paramétricos pré-definidos.

Neste capítulo, exploraremos a estrutura do modelo GAM, bem como os seus métodos de seleção de variáveis e ajuste de parâmetros.

3.2 Modelação de relações não Lineares

O Modelo Aditivo Generalizado (GAM), assim como o Modelo Linear Generalizado (GLM), assume que a variável de resposta segue uma distribuição pertencente à família exponencial. No entanto, o GAM apresenta uma vantagem em relação aos GLM, pois não exige uma relação linear entre a variável de resposta e as variáveis explicativas.

Assim, os GAM são uma extensão dos GLM, uma vez que podem incluir tanto uma componente paramétrica (soma dos preditores lineares), como uma componente não paramétrica (soma de funções suaves dos preditores lineares) [19].

3.3 Estrutura do Modelo

De um modo geral, a estrutura destes modelos é definida pela expressão:

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r + f_{r+1}(x_{r+1}) + ... + f_p(x_p) \tag{3.1}$$

onde, $\mu_i \equiv E(Y_i)$ e $x_1, ..., x_p$ são as variáveis explicativas. A variável dependente Y_i segue uma distribuição da família exponencial, como, por exemplo, a distribuição de Poisson, Gamma ou Normal. Por fim, g é uma função de ligação monótona e suave que relaciona a média da resposta, μ, com o preditor linear, $\mathbf{X}_i\beta$.

Com este modelo, é possível combinar formas paramétricas das r variáveis explicativas com os termos não paramétricos das $p - r$ variáveis. As funções suaves das variáveis explicativas são representadas por f_j, sendo estas funções contínuas que possuem derivada de todas as ordens, também estas contínuas (classe C^∞).

3.4 Funções suaves

Para simplificar a representação e estimação de funções suaves, vamos considerar um modelo que contém apenas uma função suave com uma única variável,

$$y_i = f(x_i) + \epsilon_i, \tag{3.2}$$

onde, y_i é a variável de resposta, x_i é a variável explicativa, f a função suave e ϵ_i são variáveis aleatórias independentes e identicamente distribuídas com $\epsilon_i \sim N(0, \sigma^2)$. Supomos também que $x_i \in [0, 1]$.

Para aplicar técnicas semelhantes às utilizadas no GLM, é necessário transformar o modelo 3.2 num modelo linear, o que é feito usando uma base para a função suave $f(x)$. Assim, se $b_j(x)$ for o j-ésimo elemento de uma base de comprimento m para o espaço de f, então f pode ser escrito como

$$f(x_i) = \sum_{j=1}^{m} b_j(x_i)\beta_j \tag{3.3}$$

Deste modo, combinando as equações 3.2 e 3.3, obtemos o modelo linear pretendido descrito pela seguinte equação

$$y_i = \sum_{j=1}^{m} b_j(x_i)\beta_j + \epsilon_i, \tag{3.4}$$

Assim, a questão desafiante na estimação da função suave é, por um lado, estimar os coeficientes do modelo, β_j, uma vez que pequenas variações destes valores podem resultar em curvas suaves significativamente diferentes. Na prática, o método dos mínimos quadrados penalizados iterativamente pesados (*PIRLS*) é frequentemente utilizado para essa estimação. Por outro lado, é também necessário estimar os parâmetros de suavização do modelo, que podem ser obtidos através do método de validação cruzada generalizada.

3.4.1 *Splines*

Uma *Spline* é uma curva suave formada pela união de polinómios contínuos, sendo que os pontos que unem essas secções são designados de nós [19]. Existem diferentes tipos de *splines* que podem ser definidas, sendo as *splines* cúbicas as mais comuns. Estas *splines* são compostas por secções de polinómios cúbicos, o que garante uma função contínua com derivadas de primeira e segunda ordem também contínuas em todos os pontos.

Como mencionado anteriormente, as *splines* podem ser escritas como uma combinação linear de bases de funções, da seguinte forma

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j \tag{3.5}$$

onde β_j são coeficientes constantes e $b_j(x)$ são as funções representadas por polinómios.

Para definir uma *spline* cúbica, primeiro precisamos de definir uma base. Dada a localização de n nós, denotados por $\left\{x_i^* : i = 1, ..., n\right\}$, existem várias maneiras de definir uma base para *splines* cúbicas. Neste contexto, uma abordagem simples para uma base de dimensão $n + 2$ é apresentada por Gu and Gu [20], definida por:

$$b_1(x) = 1,\ b_2(x) = x \text{ e } b_{i+2}(x) = R(x, x_i^*), \quad i = \{1, ..., n\}$$

onde,

$$\begin{aligned} R(x,z) = \quad & \left[\left(z - \frac{1}{2}\right)^2 - \frac{1}{12}\right]\left[\left(x - \frac{1}{2}\right)^2 - \frac{1}{12}\right] \times \frac{1}{4} - \\ & - \left[\left(|x - z| - \frac{1}{2}\right)^4 - \frac{1}{2}\left(|x - z| - \frac{1}{2}\right)^2 + \frac{7}{240}\right] \times \frac{1}{24} \end{aligned} \tag{3.6}$$

Após definirmos a base, os parâmetros de regressão β_j podem ser estimados pelo método dos mínimos quadrados. É importante notar que estes parâmetros dependem do número de nós e que o número e a posição dos nós têm um forte impacto na suavidade da curva final da *spline*. Assim, quantos mais nós usarmos, menos suave a curva se torna.

Portanto, se por um lado são necessários nós suficientes para capturar a tendência das observações, por outro lado, demasiados nós comprometem o grau de suavidade. Assim, a próxima questão a abordar é quantos nós devemos utilizar para controlar o grau de suavidade.

3.4.2 Como controlar o grau de suavidade

Como já foi referido, o número de nós a considerar pode afetar o grau de suavidade da curva. Deste modo, uma abordagem inicial seria tentar alcançar o número ótimo de nós utilizando testes de hipóteses. No entanto, surge um problema, uma vez que um modelo com $k-1$ nós espaçados uniformemente, normalmente não está incluído num modelo com k nós espaçados uniformemente. Portanto, uma alternativa proposta por Wood [19] para controlar o grau de suavidade é adicionar uma penalização à regressão de *splines*, em vez de focar no número ótimo de nós.

Neste caso, a suavidade do modelo é controlada pela adição de uma penalização para a oscilação no ajuste do modelo pelo método dos mínimos quadrados. Assim, em vez de minimizar apenas a soma dos quadrados dos resíduos (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \mu_i)^2 = ||y - \mu||^2 = ||\mathbf{y} - \mathbf{X}\beta||^2, \tag{3.7}$$

minimizamos a seguinte função penalizada,

$$||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \int_0^1 [f''(x)]^2 \, dx \tag{3.8}$$

A integração do quadrado da segunda derivada penaliza modelos com muitas oscilações. O equilíbrio entre o ajuste do modelo e a suavidade do modelo é controlado pelo parâmetro de suavização, λ . Quando $\lambda \to \infty$, a estimativa da função $f(x)$ aproxima-se de uma função linear, enquanto $\lambda = 0$ resulta numa *spline* de regressão não penalizada.

Como, $\int_0^1 [f''(x)]^2\,dx = \beta^T \mathbf{S}\beta$ (ex. 7, pág. 346-347 [19]), onde $\mathbf{S}$ é a matriz dos coeficientes $b_i(X)$, tal que $S_{jk} = \int b_j''(x) b_k''(x) dx$, logo é uma matriz simétrica em que as duas primeiras colunas e linhas da matriz são nulas.

Deste modo, o problema de ajuste da *spline* de regressão penalizada passa por minimizar $\mathcal{P}(\beta;\lambda)$ em ordem a β.

$$\begin{aligned} \mathcal{P}(\beta;\lambda) = &\quad ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda\beta^T\mathbf{S}\beta \\ = &\quad (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\mathbf{S}\beta \\ = &\quad \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})\beta \end{aligned} \tag{3.9}$$

Assim, o problema de estimação do grau de suavidade do modelo torna-se no problema de estimação do parâmetro de suavização λ. Mas antes de percebermos como estimar λ, vamos considerar a estimativa de β, dado λ.

Como $\mathcal{P}(\beta;\lambda)$ é uma função positiva e não limitada, o valor de $\hat{\beta}$ no qual $\mathcal{P}$ atinge o mínimo é o zero da primeira derivada, dada por

$$\begin{aligned} \frac{\partial \mathcal{P}(\beta;\lambda)}{\partial\beta} = &\quad -2(\mathbf{X}^T\mathbf{y})^T + 2\beta^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}\right) \\ = &\quad -2\mathbf{y}^T\mathbf{X} + 2\beta^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}\right) \end{aligned} \tag{3.10}$$

Assim, igualando a equação anterior a zero, vamos obter o estimador dos mínimos quadrados penalizados de β, obtido da seguinte forma:

$$\begin{aligned} &\quad \beta^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}\right) = \mathbf{y}^T\mathbf{X} \\ \Leftrightarrow &\quad \beta^T = \mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}\right)^{-1} \\ \Rightarrow &\quad \hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{3.11}$$

A vantagem deste método é que, desde que a dimensão da base seja maior do que o esperado, a escolha da base e a localização exata dos nós não vai ter uma grande influência no ajuste do modelo. No entanto, este método requer que seja escolhido um valor para λ e esta escolha pode afetar a flexibilidade do modelo e, consequentemente, a forma da curva. Assim, é importante fazer uma escolha adequada do parâmetro de suavidade, λ, algo que será explicado na próxima secção.

3.4.3 Estimação do Parâmetro de suavidade

Como já foi referido, uma escolha adequada do parâmetro de suavidade λ é importante para obter uma estimativa precisa da *spline* f, evitando que $\hat{f}$ seja bastante desviada da verdadeira função f. Se o valor de λ for muito elevado iremos obter um modelo subajustado e se for muito baixo, obtemos um modelo sobreajustado, como ilustrado na Figura 3.1.

O critério desenvolvido por Wood [19] para garantir que $\hat{f}$ esteja o mais próximo possível da sua função real é escolher um valor de λ que minimize

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2, \tag{3.12}$$

onde n é o tamanho da amostra, $\hat{f}_i \equiv \hat{f}(x_i)$ e $f_i \equiv f(x_i)$.

Como f é uma função desconhecida não é possível usar M diretamente, no entanto é possível estimar $E(M) + \sigma^2$ (erro quadrático esperado), usando validação cruzada.

A pontuação de validação cruzada ordinária (*Ordinary Cross Validation Score*) é definida como

$$v_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2 \tag{3.13}$$

onde $\hat{f}^{[-i]}$ é o modelo ajustado usando todos os dados exceto y_i.

O método consiste em excluir uma observação de cada vez, ajustar o modelo aos dados restantes e calcular o quadrado da diferença entre o valor da observação em falta e o valor previsto (Figura 3.1).

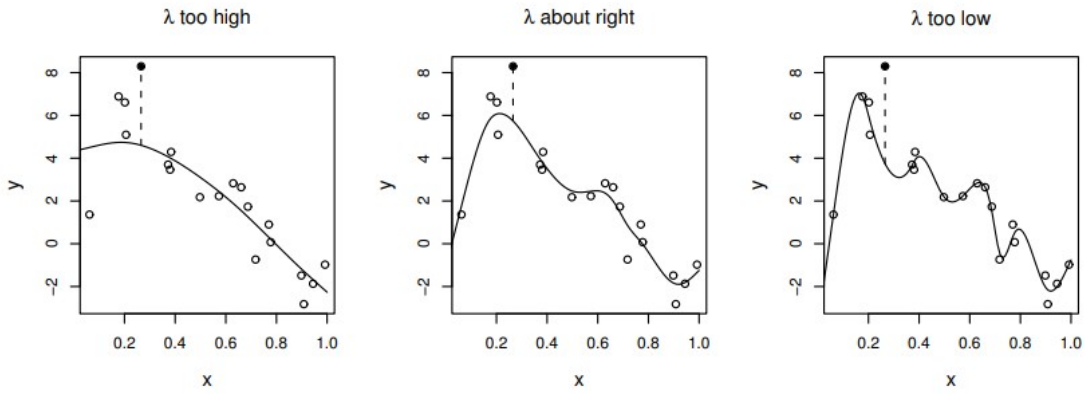


FIGURA 3.1: Ajuste da *spline* pelo método de validação cruzada. O quinto elemento (•) foi excluído do ajuste e a linha contínua é a *spline* de regressão penalizada ajustada aos restantes dados para diferentes valores de λ. (Figura adaptada de [19])

Substituindo $y_i = f_i + \epsilon_i$, obtemos

$$\begin{aligned} \nu_o = & \quad \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ = & \quad \frac{1}{n}\sum_{i=1}^{n}\left[(\hat{f}_i^{[-i]} - f_i)^2 + 2(\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2\right] \end{aligned} \tag{3.14}$$

Sendo que $\epsilon_i \sim N(0,\sigma^2) \Rightarrow \epsilon_i^2 \sim \chi(\sigma^2)$ e ϵ_i e $\hat{f}_i^{[-i]}$ são independentes, temos que o valor esperado de ν_o é dado por

$$\begin{aligned} E(\nu_o) = & \quad \frac{1}{n}E\left(\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2\right) - \frac{2}{n}E((\hat{f}_i^{[-i]} - f_i)\epsilon_i) + E(\epsilon_i^2) \\ = & \quad \frac{1}{n}E\left(\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2\right) + \sigma^2 \end{aligned} \tag{3.15}$$

Pelo Teorema do limite central podemos assumir que $\hat{f}^{[-i]} \approx \hat{f}$, e que $E(\nu_o) \approx E(M) + \sigma^2$.

Assim, uma vez que não é possível escolher λ minimizando M, é razoável escolher este parâmetro minimizando ν_o. O método de escolher λ para minimizar ν_o é conhecido como validação cruzada ordinária. No entanto, calcular ν_o tem um custo computacional elevado, uma vez que é necessário reajustar o modelo para cada um dos n conjuntos de dados resultantes. Assim, sendo

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{f}_i)^2}{(1 - A_{ii})^2} \tag{3.16}$$

onde $\hat{f}$ é a estimativa usando todos os dados e A é a matriz que relaciona os valores ajustados com os valores observados, $\hat{f} = Ay$, designada por matriz chapéu ou de influência. Na prática, A_{ii} é substituída pela média, $\frac{tr(A)}{n}$, que resulta na pontuação de validação cruzada generalizada (*Generalized Cross Validation Score*), dada pela seguinte equação

$$v_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{\left[n - tr(A)\right]^2} \tag{3.17}$$

A validação cruzada generalizada apresenta vantagens computacionais, uma vez que não requer n reajustamentos do modelo.

3.5 Estimação dos Coeficientes do Modelo

Os coeficientes do modelo são estimados pelo método dos mínimos quadrados penalizados iterativamente pesados (*Penalized Iteratively re-weighted least squares*). Neste método iteram-se os seguintes passos até atingir a convergência.

Considera-se o seguinte modelo com uma única função de suavização

$$Y_i = \beta_0 + f(X_i) + \epsilon_i, \tag{3.18}$$

onde β_0 é o termo independente, f a função suave e $\epsilon_i \sim N(0, \sigma^2)$ são as variáveis aleatórias independentes. Este modelo é estimado maximizando a função log-verosimilhança penalizada dada por

$$l_p(\beta) = l(\beta) - \frac{1}{2\phi}\lambda\beta^T\mathbf{S}\beta \tag{3.19}$$

onde λ é o parâmetro de suavização, $\mathbf{S}$ a matriz de penalização referente à função de suavização f e $l(\beta)$ é a função log-verosimilhança dada por

$$l(\beta) = \sum_{i=1}^{n} \log(f(x_i|\beta) \tag{3.20}$$

Assim, para um dado λ, os passos a realizar são:

1. Inicializar $\hat{\mu}_i = y_i + \delta_i$ e $\hat{\eta}_i = g(\hat{\mu}_i)$, sendo que δ_i normalmente é zero, no entanto pode ser uma constante pequena desde que garanta que $\hat{\eta}_i$ seja finito.

2. Dada a estimativa do preditor linear $\hat{\eta}$ e a correspondente estimativa do vetor da média da resposta, $\hat{\mu}$, calcular a pseudodata z_i e os pesos iterativos w_i:

$$w_i = \frac{1}{V(\hat{\mu}_i)g'(\hat{\mu}_i)^2} \quad \text{e} \quad z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i \tag{3.21}$$

onde var(Y_i)= $V(\mu_i)\phi$, g é uma função de ligação e ϕ um parâmetro de escala arbitrário.

3. Definindo **W** como a matriz diagonal tal que $W_{ii} = w_i$, minimizar

$$||\sqrt{\mathbf{W}}z - \sqrt{\mathbf{W}}\mathbf{X}\beta||^2 + \lambda\beta^T\mathbf{S}\beta \tag{3.22}$$

em ordem a β para obter uma nova estimativa de $\hat{\beta}$, onde **S** é a matriz de penalização, β os coeficientes do modelo e λ o parâmetro de suavização. Atualizar as estimativas $\hat{\eta} = \mathbf{X}\hat{\beta}$ e $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

3.6 Outros tipos de *splines*

É bastante comum incluir múltiplas funções suaves num modelo GAM, sendo também possível ter uma função suave aplicada a um conjunto de variáveis. Este caso é particularmente útil ao analisar dados geográficos, uma vez que é importante criar uma função suave que inclua tanto a latitude como a longitude.

A construção de *splines* pode ser entendida como um problema de otimização sujeito a restrições, cujo objetivo é minimizar a curvatura quadrática média dos pedaços polinomiais. Existem vários tipos de *splines* mencionadas por Wood [19], como as *splines* cúbicas, *splines* de regressão cúbica, *splines* de regressão cúbica cíclica, *Thin plate splines* e *Thin plate regression splines*, dependendo das restrições impostas e do grau dos pedaços polinomiais.

3.6.1 *Splines* de regressão cúbicas

Esta abordagem representa a *spline* como uma função parametrizada pelos valores que ela assume nos nós. Consideremos uma função de uma *spline* cúbica, $f(x)$, com k nós, $x_1, ..., x_k$. Seja $\beta_j = f(x_j)$ e $\delta_j = f''(x_j)$, então a *spline* pode ser escrita como

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1}, \tag{3.23}$$

se $x_j \leq x \leq x_{j+1}$ e, onde $a_j^-(x)$, $a_j^+(x)$, $c_j^-(x)$ e $c_j^+(x)$ estão definidas na Tabela 3.1 (Tabela adaptada de [19]). Wood [19] mostra que se a *spline* satisfizer as condições de ser contínua

e ter primeira e segunda derivadas também contínuas em cada nó, x_j, e a segunda derivada ser zero apenas nos nós extremos, x_1 e x_k, implica que

$$\mathbf{B}\delta^{-} = \mathbf{D}\beta \tag{3.24}$$

onde $\delta^{-} = (\delta_2, ..., \delta_{k-1})^T$ e **B** e **D** estão definidos na Tabela 3.1.

Definindo $\mathbf{F}^{-} = \mathbf{B}^{-1}\mathbf{D}$ e

$$\mathbf{F}= \begin{bmatrix} \mathbf{0} \\ \mathbf{F}^{-} \\ \mathbf{0} \end{bmatrix}$$

onde **0** é uma linha de zeros, então temos que $\delta = \mathbf{F}\beta$. Assim, podemos escrever a *spline* em termos de β da seguinte forma

$$f(x) = a_j^{-}(x)\beta_j + a_j^{+}(x)\beta_{j+1} + c_j^{-}(x)\mathbf{F}_j\beta + c_j^{+}(x)\mathbf{F}_{j+1}\beta, \tag{3.25}$$

se $x_j \leq x \leq x_{j+1}$, onde as bases de funções $a_j^{-}, a_j^{+}, c_j^{-}, c_j^{+}$ estão definidas na Tabela 3.1.

De uma forma mais condensada, a *spline* pode ser escrita como

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j \tag{3.26}$$

por definição implícita da nova base de funções $b_j(x)$.

Bases de funções para uma *spline* cúbica		
$a_j^-(x) = (x_{j+1} - x)/h_j$	$c_j^-(x) = [(x_{j+1} - x)^3/h_j - h_j(x_{j+1} - x)]/6$	
$a_j^+(x) = (x - x_j)/h_j$	$c_j^+(x) = [(x - x_j)^3/h_j - h_j(x - x_j)]/6$	
Elementos não nulos - *spline* não cíclica		
$D_{i,i} = 1/h_i$	$D_{i,i+1} = -1/h_i - 1/h_{i+1}$	$D_{i,i+2} = 1/h_{i+1}$
$B_{i,i} = (h_i + h_{i+1})/3$		$i = 1...k-2$
$B_{i,i+1} = h_{i+1}/6$	$B_{i+1,i} = h_{i+1}/6$	$i = 1...k-3$
Elementos não nulos - *spline* cíclica		
$\tilde{B}_{i-1,i} = \tilde{B}_{i,i-1} = h_{i-1}/6$	$\tilde{B}_{i,i} = (h_{i-1} + h_i)/3$	
$\tilde{D}_{i-1,i} = \tilde{D}_{i,i-1} = 1/h_{i-1}$	$\tilde{D}_{i,i} = -1/h_{i-1} - 1/h_i)$	$i = 2...k-1$
$\tilde{B}_{1,1} = (h_{k-1} + h_1)/3$	$\tilde{B}_{1,k-1} = h_{k-1}/6$	$\tilde{B}_{k-1,1} = h_{k-1}/6$
$\tilde{D}_{1,1} = -1/h_1 - 1/h_{k-1}$	$\tilde{D}_{1,k-1} = 1/h_{k-1}$	$\tilde{D}_{k-1,1} = 1/h_{k-1}$

TABELA 3.1: Definições de bases de funções e matrizes utilizadas para definir as *spline*s de regressão cúbica. Sendo que $h_j = x_{j+1} - x_j$.

3.6.2 *Splines* de regressão cúbicas cíclicas

Existem situações em que é conveniente que uma função suave seja cíclica, isto é, nos limites inferior e superior, a função tem o mesmo valor e as primeiras derivadas existem. Por exemplo, não é adequado que uma função suave da época do ano não seja contínua no final de cada ano. Para solucionar esse problema, a *spline* de regressão cúbica da secção anterior pode ser modificada para produzir estas funções suaves. A *spline* é na mesma definida pela função 3.23, mas agora temos que $\beta_1 = \beta_k$ e $\delta_1 = \delta_k$. Neste caso, definimos os vetores $\beta^T = (\beta_1, ..., \beta_{k-1})$ e $\delta^T = (\delta_1, ..., \delta_{k-1})$. As condições para que a *spline* seja contínua até à segunda derivada em cada nó e que $f(x_1)$ deve ser igual a $f(x_k)$ até à segunda derivada, são equivalentes a dizer que

$$\tilde{\mathbf{B}}\delta = \tilde{\mathbf{D}}\beta \tag{3.27}$$

onde $\tilde{\mathbf{B}}$ e $\tilde{\mathbf{D}}$ estão definidos na Tabela 3.1. De forma análoga, a *spline* pode ser escrita de forma condensada por

$$f(x) = \sum_{j=1}^{k-1} \tilde{b}_j(x)\beta_j \tag{3.28}$$

por definição da base de funções $\tilde{b}_j(x)$.

3.6.3 *Thin plate regression splines*

As bases abordadas até aqui, apesar de serem úteis na prática, sugerem alguma subjetividade no processo de ajustamento do modelo, uma vez que é necessário escolher a localização dos nós. Para além disso são também mais úteis para representar funções suaves de uma única variável explicativa e, também não está definido em que medida as bases são melhores ou piores do que outra base que possa ser usada.

Deste modo, surgem as *thin plate regression splines*, que é uma abordagem que resolve parcialmente estes problemas, produzindo bases sem nós para funções suaves com qualquer número de variáveis explicativas.

3.6.3.1 *Thin plate splines*

As *Thin plate splines* são uma solução geral para o problema de estimar a função suave com múltiplas variáveis explicativas.

Consideremos então o problema de estimar a função suave $g(x)$, a partir de n observações (y_i, x_i) tal que

$$y_i = g(x_i) + \epsilon_i$$

onde ϵ_i é o termo dos erros aleatórios e x é um vetor de dimensão d, $d \leq n$. Este método estima g encontrando a função $\hat{f}$ que minimiza

$$||\mathbf{y} - \mathbf{f}||^2 + \lambda J_{md}(f) \tag{3.29}$$

onde $\mathbf{y}$ é o vetor das observações y_i, $\mathbf{f} = (f(x_1), ..., f(x_n))^T$ e $J_{md}(f)$ é a funcional de penalização que mede as oscilações de f, m é a ordem de penalização e d é o número de variáveis explicativas. A funcional de penalização é definida por

$$J_{md}(f) = \int \cdots \int_{R^d} \sum_{\nu_1+\ldots+\nu_d=m} \frac{m!}{\nu_1!\ldots\nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}} \right)^2 dx_1 \ldots dx_d \tag{3.30}$$

Um progresso adicional só é possível se m for escolhido de modo a que $2m > d$, e pode ser mostrado que a função que minimiza 3.29 é

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md} \left(||\mathbf{x} - \mathbf{x}_i|| \right) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}), \tag{3.31}$$

onde δ e α são vetores dos coeficientes a ser estimados, estando δ sujeito às restrições lineares $T^T\delta = 0$, onde $T_{ij} = \phi_j(\mathbf{x}_i)$. As $M = \binom{m+d-1}{d}$ funções ϕ_i são polinómios linearmente independentes abrangendo o espaço de polinómios em R^d de grau menor ou igual a m, esta função abrange o espaço de funções tais que $J_{md}(f) = 0$, ou seja o espaço nulo de $J_{md}(f)$.

Por exemplo, se $m = d = 2$ temos os polinómios de R^2, 2 variáveis, com grau inferior a 2. Estas funções são $\phi_1(x) = 1$, $\phi_2(x) = x_1$ e $\phi_3(x) = x_2$. As restantes bases de funções usadas em 3.31 são definidas por

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d} \log(r) & \text{se } d \text{ é par} \\ \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & \text{se } d \text{ é ímpar} \end{cases}$$

Definindo $\mathbf{E}$ por $E_{ij} \equiv \eta_{md}(||\mathbf{x}_i - \mathbf{x}_j)||$, então o problema de ajustamento da *Thin plate spline* passa a ser minimizar a função 3.32 em ordem a δ e α

$$||\mathbf{y} - \mathbf{E}\delta - \mathbf{T}\alpha||^2 + \lambda\delta^T\mathbf{E}\delta \tag{3.32}$$

restrito a $\mathbf{T}^T\delta = 0$.

À primeira vista, parece que todos os problemas listados no início desta secção foram resolvidos, uma vez que não foi necessário escolher as posições dos nós nem selecionar

funções de base, o que leva a crer que as *Thin plate splines* devem ser usadas para representar todos os termos suaves no modelo.

No entanto, surge um problema com estas *splines*, uma vez que têm um custo computacional bastante elevado, pois têm tantos parâmetros desconhecidos como elementos dos dados. O custo computacional da estimativa do modelo é proporcional ao cubo do número de parâmetros, exceto no caso de existir apenas uma variável explicativa.

3.6.3.2 *Thin plate regression splines*

As *Thin plate regression splines* são baseadas na ideia de truncamento do espaço das componentes onduladas das *Thin plate splines* (componentes com parâmetros δ) e mantendo as componentes com ondulação nula inalteradas (componentes com parâmetros α), com o objetivo de melhorar o desempenho computacional das *Thin plate splines*.
Seja $\mathbf{UDU}^T$ a decomposição em valores e vetores próprios de $\mathbf{E}$, onde $\mathbf{D}$ é a matriz diagonal dos valores próprios, tal que $|D_{i,i}| \geq |D_{i-1,i-1}|$ e as colunas de $\mathbf{U}$ são os vetores próprios correspondentes.

Sendo $\mathbf{U}_k$ a matriz formada pelas primeiras k colunas de $\mathbf{U}$ e $\mathbf{D}_k$ a submatriz superior esquerda $k \times k$ de $\mathbf{D}$. Restringindo δ ao espaço de colunas de $\mathbf{U}_k$, escrevendo $\delta = \mathbf{U}_k\delta_k$, significa que a equação a minimizar, em ordem a δ_k e α é agora escrita como

$$||\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\delta_k - \mathbf{T}\alpha||^2 + \lambda\delta_k^T\mathbf{D}_k\delta_k \tag{3.33}$$

restrito a $\mathbf{T}^T\mathbf{U}_k\delta_k = 0$.

Em primeiro lugar, encontramos uma base em que os vetores coluna da matriz são ortogonais, $\mathbf{Z}_k$, em que $\mathbf{T}^T\mathbf{U}_k\mathbf{Z}_k = \mathbf{0}$. Uma maneira de o fazer é usando a decomposição $\mathbf{QR}$ ($\mathbf{Q}$ é uma matriz ortogonal e $\mathbf{R}$ uma matriz triangular superior) de $\mathbf{T}^T\mathbf{Z}_k$ [19]. Restringindo δ_k a este espaço, escrevendo $\delta_k = \mathbf{Z}_k\tilde{\delta}$, resulta no problema sem restrições que deve ser resolvido para ajustar a aproximação de ordem k da *spline*, minimizando 3.34 em ordem a $\tilde{\delta}$ e α

$$||\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\mathbf{Z}_k\tilde{\delta} - \mathbf{T}\alpha||^2 + \lambda\tilde{\delta}^T\mathbf{Z}_k^T\mathbf{D}_k\mathbf{Z}_k\tilde{\delta} \tag{3.34}$$

Este problema tem um custo computacional de $O(k^3)$. No entanto, o principal problema é como encontrar as matrizes $\mathbf{U}_k$ e $\mathbf{D}_k$ de uma maneira computacionalmente eficiente. Uma decomposição completa de valores próprios de $\mathbf{E}$ requer $O(n^3)$ operações, o que limitaria a utilidade da abordagem das *Thin plate regression splines*. Felizmente, o método de iteração de Lanczos pode ser empregado para encontrar estas matrizes com um custo substancialmente menor, de $O(n^2k)$ operações (ver apêndice A, secção 11 Wood [19]).

Capítulo 4

Aplicação a uma base de dados

Este capítulo faz uma descrição detalhada da organização da base de dados, seguida de uma análise exploratória e, por fim, a aplicação prática dos métodos abordados nos capítulos anteriores, isto é, do processo de imputação múltipla e da aplicação de modelos e interpretação dos mesmos.

4.1 Organização da base de dados

O estudo utilizou a base de dados do Sistema Nacional de Vigilância da Tuberculose em Portugal (SVIG-TB). A base de dados utilizada contém 15.359 observações, referentes a todos os pacientes diagnosticados com tuberculose com residência nos diversos municípios de Portugal notificados entre o período de 2008 a 2017 no SVIG-TB (Tabela 4.1).

A variável de resposta em estudo é o tempo de atraso do diagnóstico da tuberculose nos cuidados de saúde, referida como $delay\_health$. O valor da variável corresponde à diferença entre a data da primeira consulta e a data do diagnóstico. Foram excluídos do estudo os pacientes que apresentavam dados omissos ou incorretos para as datas da primeira consulta e/ou do diagnóstico, uma vez que isso impossibilitaria o cálculo do atraso nos cuidados de saúde. Além disso, foram também excluídos pacientes que apresentaram um atraso total de mais de um ano (> 365 dias) entre o início dos sintomas e o diagnóstico, uma vez que é altamente improvável que um atraso tão longo ocorra, sugerindo a possibilidade de erros na introdução das datas que definem o tempo até o diagnóstico.

Das 69 variáveis disponíveis na base de dados, apenas 35 são consideradas relevantes segundo profissionais da área. Devido ao elevado número de variáveis, estas estão descritas no Apêndice A, nas Tabelas A.1 e A.2. Falta apenas referir que, observações

referentes a indivíduos residentes nas ilhas dos Açores e da Madeira não foram tidas em consideração, cingindo assim o estudo a Portugal Continental. Após esta filtragem, a base de dados resultante contém 12.882 observações.

Ano	N	Tempo de atraso mediano nos CS
2008	1646	9
2009	1594	8
2010	1471	8
2011	1453	8
2012	1362	8
2013	1240	10
2014	1140	9
2015	1072	9
2016	950	9
2017	954	10

TABELA 4.1: Tempo de atraso mediano nos cuidados de saúde por ano

Através da Tabela 4.1 observamos que o tempo de atraso mediano nos cuidados de saúde mantém-se aproximadamente constante ao longo dos anos.

Apontadas as principais características do conjunto de dados, passamos agora à análise detalhada dos mesmos.

4.2 Análise exploratória dos dados

Para entender como os dados se comportam é importante fazer uma análise descritiva, gráfica e numérica das variáveis.

O conjunto de dados contém um total de 4 variáveis contínuas, escolaridade, casa, beneficiários e médicos. A seguir, estão apresentados os histogramas para cada variável (Figura 4.1), através dos quais podemos fazer observações sobre a simetria/assimetria de cada variável, bem como a presença de outliers.

Analisando a Figura 4.1, consideramos como variáveis simétricas as variáveis escolaridade, médicos e casa, apesar desta última apresentar alguns outliers. Para este caso são fornecidas a média e o desvio padrão, uma vez que apresentam um comportamento próximo da distribuição normal. Já para a variável beneficiários, a única variável assimétrica, são apresentados os valores da mediana, mínimo e máximo. Estes resultados estão apresentados na Tabela 4.2. Para além disso, verifica-se que nenhuma das variáveis contínuas apresenta valores omissos.

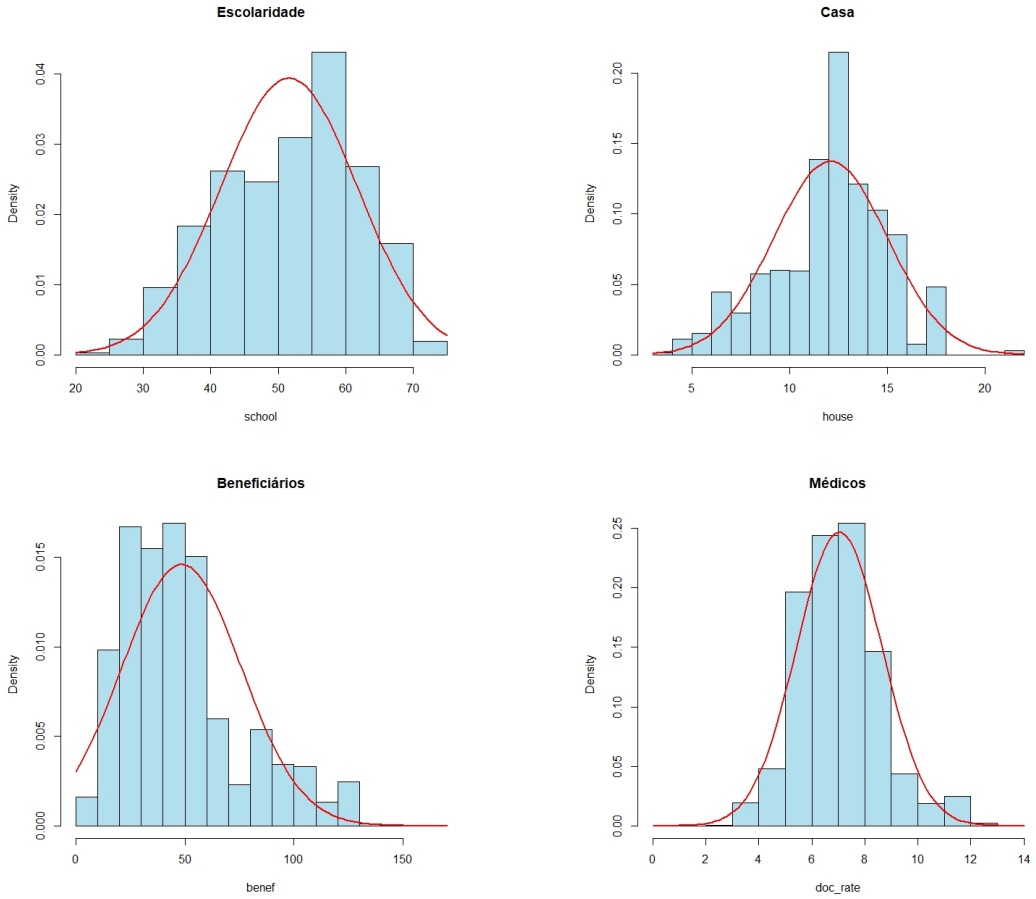


FIGURA 4.1: Histogramas das variáveis contínuas

Variável simétrica	Média	Desvio padrão	Valores omissos
Escolaridade	51.62	10.13	0.00%
Casa	12.11	2.91	0.00%
Médicos	7.04	1.62	0.00%
Variável assimétrica	**Mediana**	**Mínimo - Máximo**	**Valores omissos**
Beneficiários	43.12	3.48 - 164.33	0.00%

TABELA 4.2: Descrição numérica - variáveis contínuas (N = 12 882)

Outro aspeto a ser analisado é a correlação entre as variáveis para avaliar se as variáveis explicativas são independentes entre si. Para avaliar a correlação entre as variáveis contínuas, optou-se pelo coeficiente de correlação de *Spearman*, esta escolha deve-se ao facto de uma das variáveis não apresentar uma distribuição próxima da normal.

A Figura 4.2 apresenta a matriz de correlação das variáveis contínuas. Podemos observar que nenhum par de variáveis está altamente correlacionado, sendo que as duas mais correlacionadas são as variáveis casa e beneficiários, com um coeficiente de correlação de 0.38.

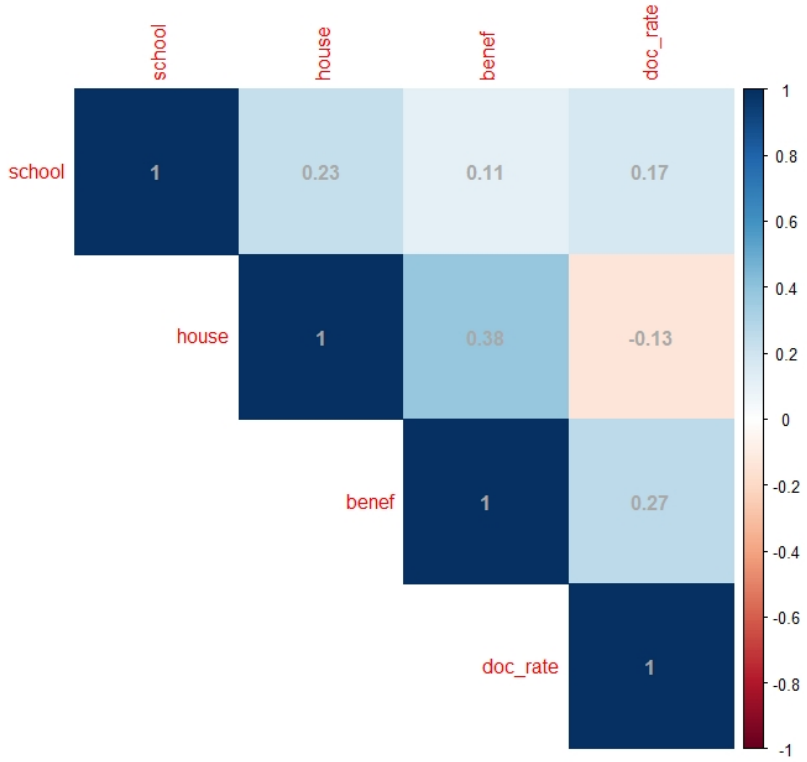


FIGURA 4.2: Cronograma dos coeficientes de correlação de Spearman

O conjunto de dados contém 30 variáveis categóricas, das quais 23 são binárias e as restantes possuem mais do que duas categorias. Através da Tabela descritiva (B.1) presente no Apêndice B é possível analisar a distribuição das observações em cada variável através da frequência absoluta e relativa. Podemos observar que, aproximadamente 70% dos pacientes são do sexo masculino e 45% têm entre 20 a 44 anos de idade.

De seguida, são apresentados os boxplots (Figura 4.3 e 4.4) para cada categoria de algumas variáveis categóricas, com o objetivo de compreender como a variável resposta se comporta nas diferentes categorias. Para facilitar a visualização dos gráficos, foi aplicado o logaritmo da variável resposta.

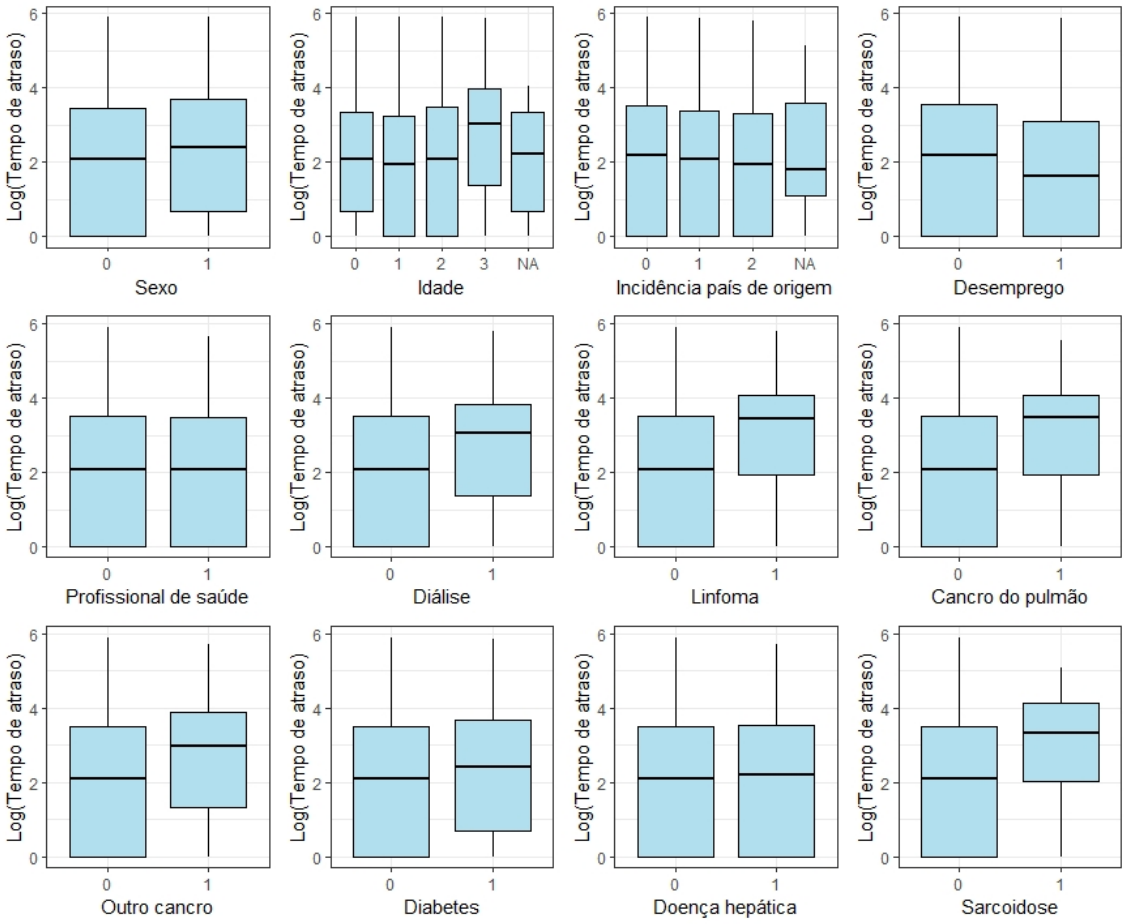


FIGURA 4.3: Boxplots variáveis categóricas - 1º grupo

As Figuras 4.3 e 4.4, bem como a Tabela B.1 permitem observar que ocorre um atraso superior em pacientes com as seguintes caracteristicas: sexo masculino, faixa etária $\geq$ 65 anos, empregados, com insuficiência renal (diálise), linfoma, cancro do pulmão, outro cancro, diabetes, sarcoidose, doença articular inflamatória, DPOC, silicose, intersticial, outra doença e ser recluso. Por outro lado, verifica-se um menor atraso em pacientes consumidores de álcool e drogas, que vivem na rua ou que vivem em residências comunitárias.

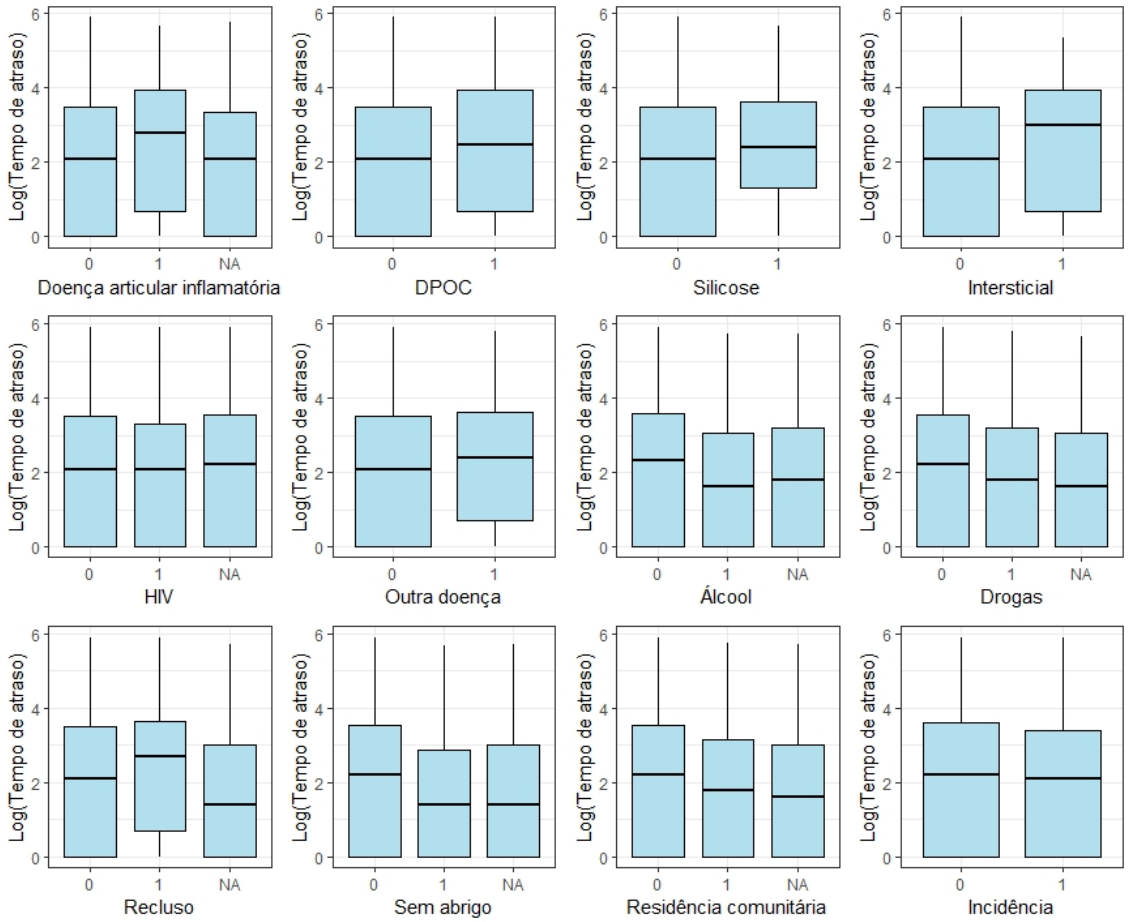


FIGURA 4.4: Boxplots variáveis categóricas - 2º grupo

Como mencionado anteriormente, a variável resposta é o tempo de atraso do diagnóstico da tuberculose nos cuidados de saúde, expresso em dias. Para os casos em que o tempo de atraso é zero, foi atribuído o valor de um dia para permitir a aplicação do logaritmo aos dados, na tentativa de os aproximar o mais possível de uma distribuição normal.

A Figura 4.5 apresenta o histograma do tempo de atraso nos cuidados de saúde antes e depois da aplicação do logaritmo. A Tabela 4.3 mostra as medidas descritivas da variável resposta, tanto para os dados originais como para os dados transformados pelo logaritmo.

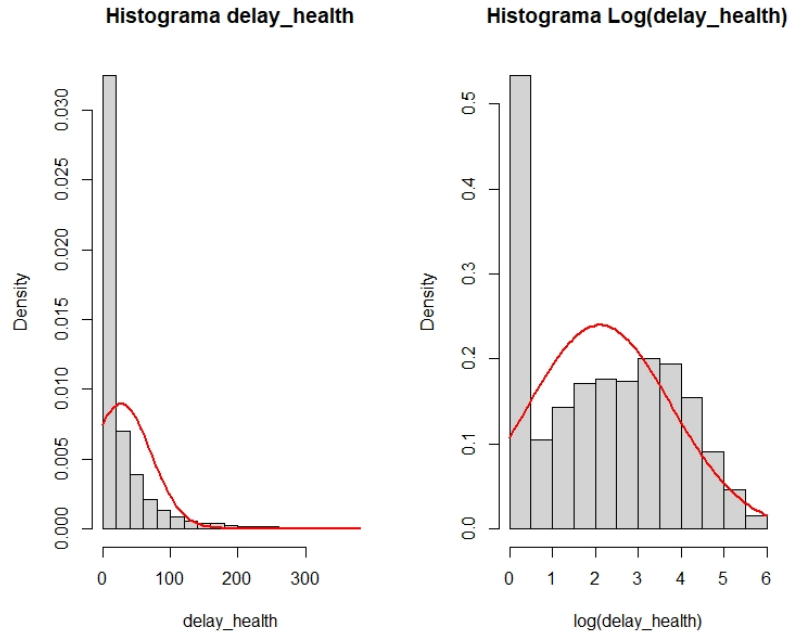


FIGURA 4.5: Histograma do tempo de atraso nos cuidados de saúde antes (esquerda) e depois (direita) de aplicar o logaritmo

	Tempo de atraso nos cuidados de saúde	
	Sem Transformação	Transformação Logarítmica
Mínimo	1	0
Mediana	8	2.08
Média	27.13	2.11
Máximo	365	5.90
Desvio padrão	44.51	1.66

TABELA 4.3: Descrição numérica da variável resposta

Ao analisar a Figura 4.5 e a Tabela 4.3, constatamos que, mesmo depois de aplicada a transformação logarítmica, a distribuição do tempo de atraso não é simétrica. Isto sugere que a distribuição gaussiana pode não ser a mais adequada para descrever os dados transformados.

Portanto, para determinar qual a distribuição da família exponencial que melhor se ajusta à variável de resposta, foram utilizadas as funções `fitdist`, `denscomp`, `cdfcomp`, `qqcomp` e `gofstat` do pacote `fitdistrplus` do `R`. Estas funções permitem ajustar diferentes distribuições aos dados e realizar uma análise comparativa utilizando gráficos e estatísticas de ajuste.

O gráfico quantil-quantil presente na Figura 4.6 compara os quantis teóricos das três distribuições ajustadas com os quantis empíricos dos dados.

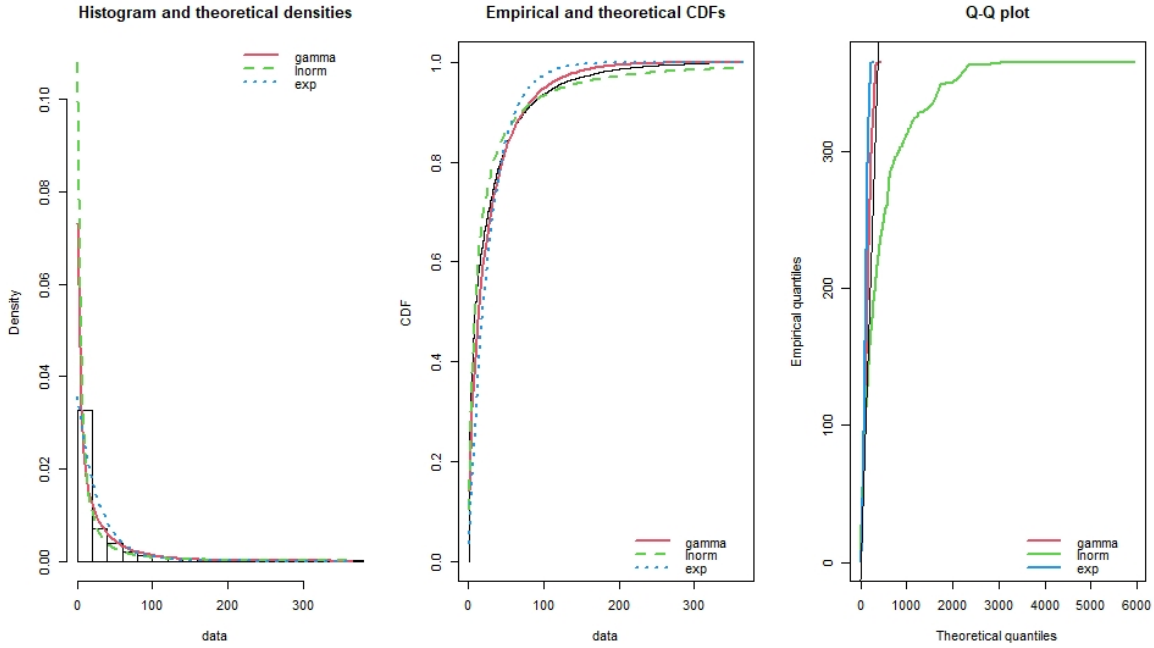


FIGURA 4.6: À esquerda o histograma dos dados em relação às funções de densidade ajustadas (gamma, lognormal e exponencial); ao centro a função de distribuição cumulativa empírica em relação às funções de distribuição ajustados; e gráfico quantil-quantil, à direita.

Estatística	**Gamma**	**Lnorm**	**Exp**
Kolmogorov-Smirnov	0.14	0.16	0.25
Cramer-von Mises	55.27	42.21	293.22
Anderson-Darling	365.68	335.56	2186.05

TABELA 4.4: Estatísticas de ajustamento aos dados

A Tabela 4.4 inclui as estatísticas de ajuste aos dados, em particular a estatística de Kolmogorov-Smirnov, a estatística de Cramer-von Mises e a estatística de Anderson-Darling. Quanto menor o valor destas estatísticas, melhor o ajuste da distribuição teórica aos dados.

Com base nos resultados acima apresentados, em especial graficamente (Figura 4.6), a distribuição gamma foi a escolhida, mesmo que dois dos critérios não estejam de acordo com esta escolha.

Terminada a análise exploratória, seguimos para a implementação do método de imputação múltipla.

4.3 Imputação

Como já foi referido, antes de iniciar o processo de imputação múltipla é necessário explorar e perceber o padrão dos valores omissos.

A Tabela 4.5 mostra a contagem e a percentagem de valores omissos em cada variável.

Variável	Total de NA's	Percentagem de NA's
age_group	10	0.08 %
country_inc	8	0.06 %
inflam_joint	233	1.78 %
HIV	1092	8.33 %
alcohol	578	4.41 %
drugs	477	3.64 %
inmate	290	2.21 %
homeless	295	2.25 %
commu_resid	336	2.56 %

TABELA 4.5: Contagem de valores omissos em cada variável

Verificamos que, num total de 35 variáveis, existem 26 completas, ou seja sem nenhum valor omisso, e, que das variáveis incompletas nenhuma tem uma percentagem de NA's muito elevada, o que é indicador de uma forte probabilidade de sucesso na implementação do método de imputação múltipla.

Uma suposição do método de imputação múltipla é que os dados sejam *Missing at Random* (MAR), ou seja que a probabilidade de um valor estar ausente pode depender de outras variáveis observadas, mas não depende do próprio valor ausente. Se esta suposição não for válida, a imputação múltipla pode levar a estimativas enviesadas.

Para criar as imputações utilizamos a função `mice()` do pacote `mice` do `R`. Esta função pressupõe uma distribuição para cada variável e faz a imputação dos valores ausentes de acordo com essa distribuição. O `mice` escolhe automaticamente as distribuições para cada variável, neste caso como só temos variáveis categóricas a serem imputadas foram apenas utilizadas a regressão logística (*logreg*) e a regressão logística politómica (*polyreg*).

A regressão logística é utilizada quando estamos perante uma variável resposta binária. As variáveis que utilizaram este método foram *inflam_joint*, *HIV*, *alcohol*, *drugs*, *inmate*, *homeless* e *commu_resid*. A regressão logística politómica, também conhecida como regressão logística multinomial, é um método de classificação que generaliza a regressão logística para variáveis com mais de dois resultados discretos possíveis. Foi utilizado

para preencher a variável de quatro categorias *age_group* e a variável de três categorias *country_inc*.

O resultado final do `mice` está apresentado na Figura 4.7 e foi obtido através do seguinte código, onde m é o número de imputações (neste caso iremos utilizar $m = 5$) e o uso de `seed` garante a reprodutibilidade das imputações.

```
imp <- mice(dadosf, m = 5, seed = 1)
```

```
Class: mids
Number of multiple imputations:  5
Imputation methods:
          sex     age_group   country_inc  unemployment    health_job      dialysis      lymphoma   lung_cancer
           ""     "polyreg"     "polyreg"            ""            ""            ""            ""            ""
 other_cancer      diabetes         liver   sarcoidosis  inflam_joint          COPD     silicosis  interstitial
           ""            ""            ""            ""      "logreg"            ""            ""            ""
          HIV other_disease       alcohol         drugs        inmate      homeless   commu_resid  delay_health
    "logreg"            ""      "logreg"      "logreg"      "logreg"      "logreg"      "logreg"            ""
      school         house         benef      doc_rate       inc_cat             n             x             y
          ""            ""            ""            ""            ""            ""            ""            ""
PredictorMatrix:
             sex age_group country_inc unemployment health_job dialysis lymphoma lung_cancer other_cancer diabetes liver
sex            0         1           1            1          1        1        1           1            1        1     1
age_group      1         0           1            1          1        1        1           1            1        1     1
country_inc    1         1           0            1          1        1        1           1            1        1     1
unemployment   1         1           1            0          1        1        1           1            1        1     1
health_job     1         1           1            1          0        1        1           1            1        1     1
dialysis       1         1           1            1          1        0        1           1            1        1     1
             sarcoidosis inflam_joint COPD silicosis interstitial HIV other_disease alcohol drugs inmate homeless
sex                    1           1    1         1            1   1             1       1     1      1        1
age_group              1           1    1         1            1   1             1       1     1      1        1
country_inc            1           1    1         1            1   1             1       1     1      1        1
unemployment           1           1    1         1            1   1             1       1     1      1        1
health_job             1           1    1         1            1   1             1       1     1      1        1
dialysis               1           1    1         1            1   1             1       1     1      1        1
             commu_resid delay_health school house benef doc_rate inc_cat n x y
sex                    1            1      1     1     1        1       1 1 1 1
age_group              1            1      1     1     1        1       1 1 1 1
country_inc            1            1      1     1     1        1       1 1 1 1
unemployment           1            1      1     1     1        1       1 1 1 1
health_job             1            1      1     1     1        1       1 1 1 1
dialysis               1            1      1     1     1        1       1 1 1 1
```

FIGURA 4.7: Output `mice`

O conjunto de dados com múltiplas imputações é armazenado no objeto `imp` de classe `mids`. A matriz de predição informa-nos que variáveis irão ser utilizadas para prever um valor plausível para outras variáveis (1 indica que uma variável é utilizada para prever outra, 0 caso contrário). Como nenhuma variável pode prever-se a si mesma, a interseção de uma variável consigo mesma na matriz recebe o valor 0. Neste caso, devido ao elevado número de variáveis explicativas presentes no conjunto de dados, excluem-se da previsão as variáveis *DICOMUN*, *district*, *county*, *year*, *region* e *problem_health*.

Após os valores serem imputados é importante inspecionar as distribuições dos dados originais e imputados, com recurso à função `stripplot()`.

A Figura 4.8 mostra as distribuições das nove variáveis que foram imputadas como pontos individuais, os pontos azuis são os valores observados, enquanto os pontos vermelhos são os valores imputados. A semelhança no posicionamento dos pontos vermelhos e azuis indica que os valores imputados aparentam ser valores plausíveis.

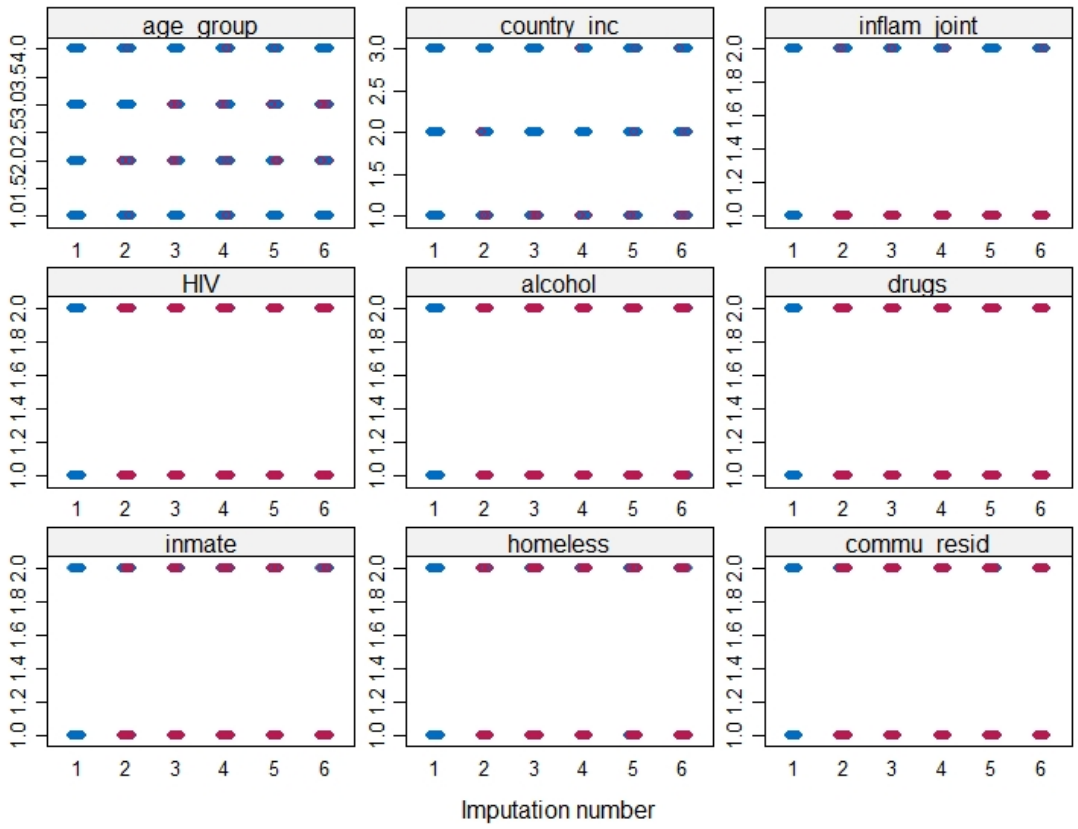


FIGURA 4.8: Stripplot das variáveis nos dados originais e nos cinco conjuntos de dados imputados.

Os próximos passos da imputação múltipla, análise e combinação, vão ser mostrados na secção 4.4.3 com o modelo final. As duas etapas foram obtidas através das funções `with()` e `pool()`, respetivamente do pacote `mice`. Todo o processo de modelação está descrito na secção seguinte.

4.4 Modelação

Em primeiro lugar, teremos que fazer a análise da distribuição geográfica das observações de modo a perceber se a variável espacial será ou não incluída no modelo final.

4.4.1 Análise Espacial

Foram adicionadas três variáveis ao conjunto de dados para que fosse possível realizar uma análise espacial dos mesmos. As variáveis acrescentadas referem-se à longitude e

latitude do centróide de cada município e uma última referente ao número de observações em cada município.

Inicialmente, analisamos o efeito bruto das variáveis espaciais na variável resposta. Nesta fase, não foi necessário utilizar os dados imputados, uma vez que todas as variáveis utilizadas para este modelo estão completas. Assim, utilizamos a função `gam()` do pacote `mgcv`, por ser um método eficaz para modelar relações não lineares complexas, como já foi referido, o que é especialmente relevante em análises espaciais. O modelo foi implementado da seguinte forma:

```
m1 <- gam(delay_health ~ s(x,y,k=100,bs='tp') + offset(log(n)),
      data = dados_merged, family = Gamma(log))
```

Onde s representa a função suave, neste caso do tipo thin plate, x e y são as variáveis longitude e a latitude, respetivamente e n é o número de observações de cada município.

Incorporámos o termo $offset(log(n))$ ao modelo para ajustar as estimativas tendo em conta o número de observações de cada município.

	Estimativa	**Std. Error**	**t value**	$Pr(>\|t\|)$
(Intercept)	-1.603	0.018	-89.22	$< 2e-16$
	edf	**Ref.df**	**F**	**valor-p**
s(x,y)	93.52	98.19	62.35	$< 2e-16$

R-sq.(adj)	= -0.244	Desviância explicada	= 45.7%	
GCV	= 2.6297	Scale est.	= 4.1614	n = 12882

TABELA 4.6: Sumário do modelo considerando apenas a variável espacial

Na Tabela 4.6 está apresentado um resumo do sumário do modelo considerando apenas o efeito espacial, podemos verificar que o termo suave é significativo (valor p $<$ 2e-16), o que indica que a localização geográfica (x e y) tem um efeito significativo no atraso do diagnóstico nos cuidados de saúde. O número de graus de liberdade efetivos para o termo suave é de 93.52, o que indica que o efeito espacial é bastante complexo. A desviância explicada é de 45.7%, ou seja cerca de 46% da variabilidade dos dados pode ser explicada pelo modelo. No entanto, este é um modelo bruto e, por isso esta medida pode aumentar quando adicionamos mais variáveis explicativas ao modelo.

De seguida, vamos analisar a distribuição geográfica do atraso do diagnóstico nos cuidados de saúde (Figura 4.9).

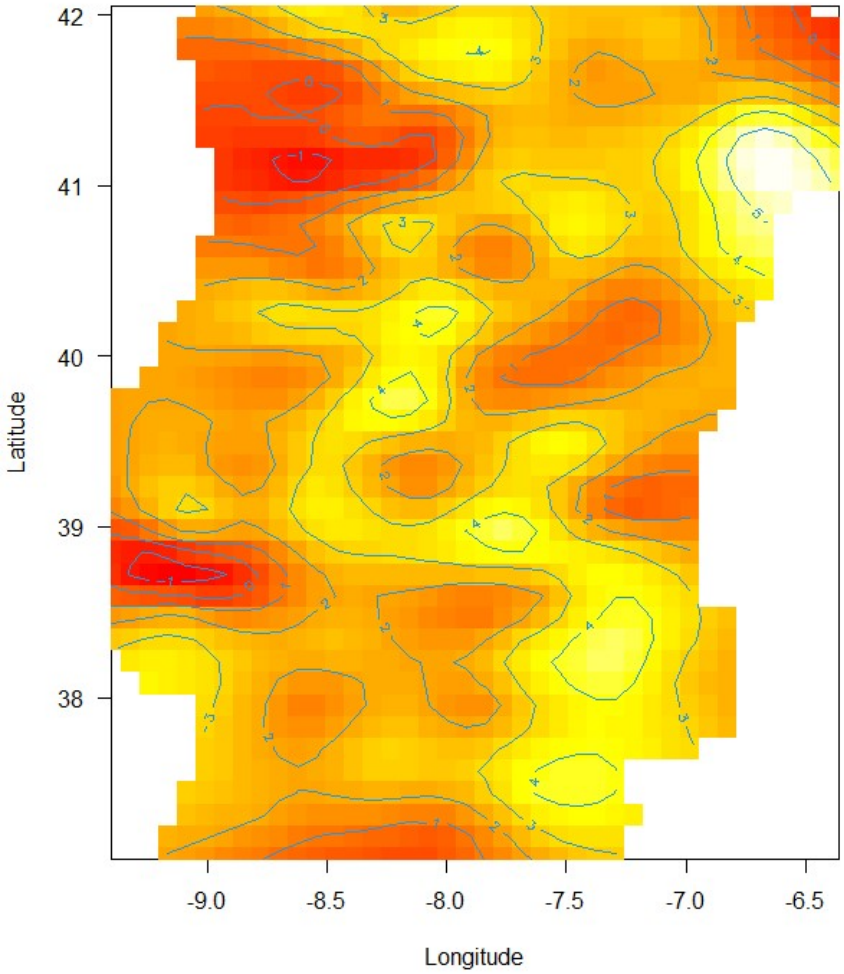


FIGURA 4.9: Distribuição geográfica do atraso do diagnóstico nos cuidados de saúde

O modelo representado na Figura 4.9 não apresenta a distribuição real do atraso do diagnóstico, mas sim a variação prevista deste atraso tendo em conta apenas a distribuição geográfica. As áreas a amarelo representam locais com maior atraso do diagnóstico nos cuidados de saúde, as áreas a vermelho representam os locais com menor atraso médio. Por fim, as linhas verdes representam as curvas de nível em escala logarítmica que comparam o atraso do diagnóstico local com o valor médio do atraso considerado e Portugal continental.

A partir desta análise, concluímos que a componente espacial deve ser incluída no modelo final para termos uma representação mais precisa do atraso do diagnóstico nos cuidados de saúde. Esta inclusão resulta diretamente do valor-p observado (Tabela 4.6).

4.4.2 Variáveis contínuas

Para compreender a relação entre cada variável explicativa contínua e a variável resposta, analisamos graficamente essa relação através da Figura 4.10. Estes gráficos fornecem-nos uma ideia inicial sobre possíveis tendências e relações não lineares.

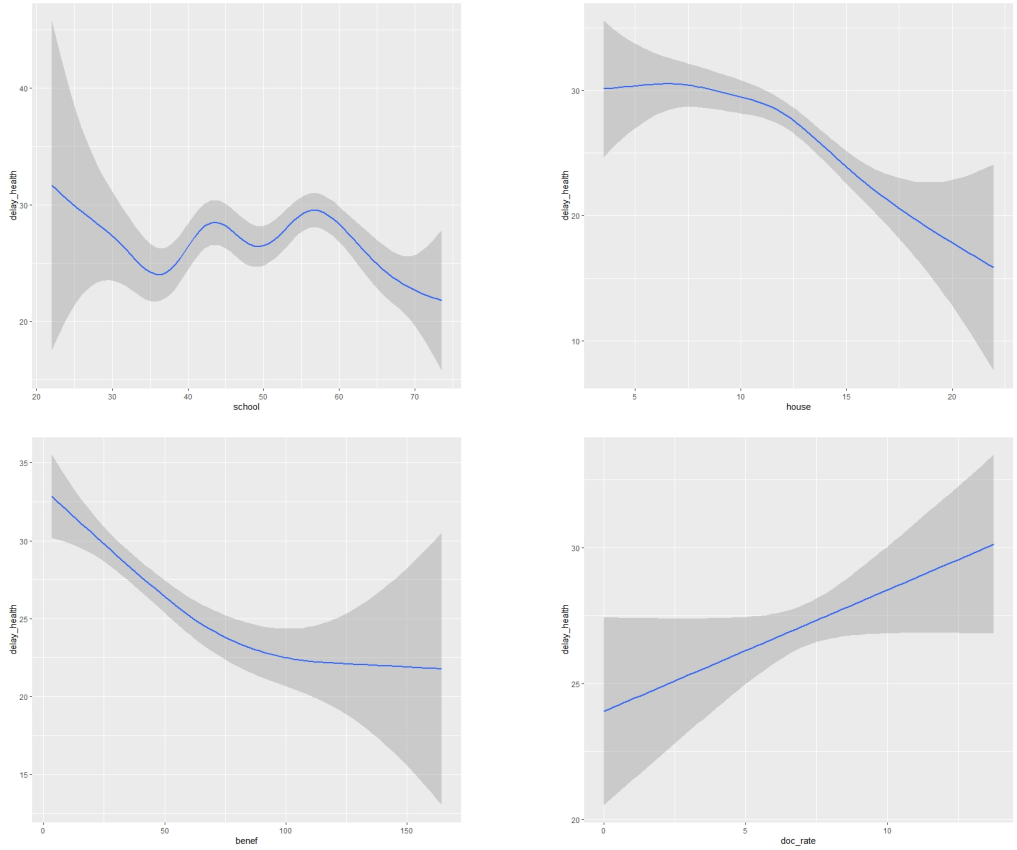


FIGURA 4.10: Relação entre variáveis explicativas contínuas e variável resposta

Observamos que existe uma tendência linear decrescente nas variáveis *house* e *benef*, e uma tendência linear crescente na variável *doc_rate*. Note-se que, o cariz linear crescente da relação entre a variável *doc_rate* e a resposta sugere uma associação não intuitiva das duas. Interpretamos do gráfico que a um aumento na proporção de médicos por 10 000 indivíduos (o que implica um aumento na quantidade de médicos) está associado o aumento do atraso no diagnóstico nos cuidados de saúde. Ora, isto à primeira vista não faz sentido corresponder à realidade, uma vez que, à partida, quantos mais médicos disponíveis, menor o atraso.

No Apêndice C, as Tabelas C.1 e C.2 dão conta dos valores de correlação entre a variável *doc_rate* e as variáveis categóricas, numa tentativa de justificar a relação sugerida pelo gráfico, no entanto, estes mesmos valores, sendo no geral residuais, não contribuem

para esta mesma justificação. Sendo assim, e uma vez que no ajuste dos modelos univariados e do modelo final, a variável *doc_rate* mostra-se, invariavelmente, não significativa para um nível de significância de 0.05, conclui-se que qualquer informação retirada do gráfico *doc_rate* vs resposta, Figura 4.10, não deve ser tomada como verdadeira.

De seguida, para avaliar quantitativamente a influência de cada variável, aplicamos modelos univariados paramétricos e não paramétricos.

Modelo univariado paramétrico :

$log(E(delay\_health)) = \beta_0 + \beta_1 \text{variável} + \epsilon$, onde $delay\_health \sim \Gamma(\alpha, \beta)$

$E(delay\_health) = \mu = \frac{\alpha}{\beta}$, $Var(delay\_health) = \frac{\alpha}{\beta^2}$ e $g(\mu) = log(\mu)$

Implementação do modelo em `R` :

```
mod_param <- gam(delay_health ~ variável,
                 data = dados_merged, family = Gamma(log))
```

Variável	Coeficiente	Valor-p	Desviância Explicada (%)	R-sq.(adj)
school	8.71e-05	0.95	3.1e-05	-7.74e-05
house	-0.034	5.1e-12	0.398	0.0031
benef	-0.0036	1.08e-11	0.41	0.0036
doc_rate	0.016	0.068	0.0295	0.00019

TABELA 4.7: Sumário modelos univariados paramétricos

Modelo univariado não paramétrico :

$log(E(delay\_health)) = \beta_0 + \beta_1\, f(\text{variável}) + \epsilon$, onde $delay\_health \sim \Gamma(\alpha, \beta)$

$E(delay\_health) = \mu = \frac{\alpha}{\beta}$, $Var(delay\_health) = \frac{\alpha}{\beta^2}$ e $g(\mu) = log(\mu)$

Implementação do modelo em `R` :

```
mod_n_param <- gam(delay_health ~ s(variável),
                   data = dados_merged, family = Gamma(log))
```

Variável	Graus de Liberdade	Valor-p	Desviância Explicada (%)	R-sq.(adj)
school	7.77	2.31e-05	0.345	0.0024
house	5.588	<2e-16	0.555	0.0042
benef	2.222	<2e-16	0.456	0.0038
doc_rate	3.427	0.309	0.0568	0.00024

TABELA 4.8: Sumário modelos univariados não paramétricos

A partir da Tabela 4.7, observamos que no modelo univariado paramétrico, a variável *house* possui um valor-p muito baixo (5.1e-12), sugerindo uma influência estatisticamente significativa. Da mesma forma, a variável *benef* tem um valor-p também muito pequeno (1.08e-11). A variável *doc_rate* apresenta uma influência positiva, com um coeficiente de 0.016. A não significância desta variável, já mencionada anteriormente, verifica-se na Tabela 4.7 através de um valor-p superior a 0.05.

Por outro lado, no modelo univariado não paramétrico (Tabela 4.8), a variável *house* continua a ter uma importância significativa, mas agora com uma maior percentagem de desviância explicada (0.555 %). Todas as variáveis, exceto *doc_rate*, apresentam um valor-p extremamente baixo, implicando alta significância estatística.

Sabendo que o modelo paramétrico é mais apropriado para interpretar a influência das variáveis e, considerando que, em geral, as conclusões de ambos os modelos são semelhantes, o modelo paramétrico é preferível por oferecer uma interpretação mais clara dos efeitos na variável resposta.

Desta forma, as quatro variáveis contínuas serão incluídas no modelo final como termos paramétricos.

4.4.3 Construção do modelo final

Para a construção do modelo final, utilizamos novamente a função `gam` do pacote `mgcv` do `R`. O modelo inicial foi construído de modo a incluir todas as variáveis disponíveis como termos paramétricos, à exceção das variáveis longitude e latitude de cada município, que foram representadas por um termo não paramétrico.

Utilizamos o método de seleção de variáveis *backward elimination*. Este método começa com um modelo completo que inclui todas as variáveis consideradas e, de seguida, uma a uma, as variáveis menos significativas são removidas, começando com aquela que tem o maior valor-p.

Para a variável explicativa sexo, escolhemos como categoria de referência o sexo masculino, o grupo escolhido para categoria de referência para a variável da idade foram os pacientes com idade compreendida entre os 20 e 44 anos. Para a variável incidência a categoria de referência é a incidência baixa.

O modelo final inclui as variáveis sexo, idade, consumo abusivo de álcool, recluso, residência comunitária, escolaridade, casa, beneficiários, incidência e, ainda, as coordenadas geográficas como termo não paramétrico, todas estas variáveis mostraram-se estatisticamente significativas.

Modelo final :

$log(E(delay\_health)) = \beta_0 + \beta_1 sex + \beta_2 age\_group0 + \beta_3 age\_group2 + \beta_4 age\_group3 + \beta_5 alcohol + \beta_6 inmate + \beta_7 commu\_resid + \beta_8 school + \beta_9 house + \beta_{10} benef + \beta_{11} inc\_cat + \beta_{12} f(x,y) + \epsilon$, onde $delay\_health \sim \Gamma(\alpha, \beta)$

$E(delay\_health) = \mu = \frac{\alpha}{\beta}$, $Var(delay\_health) = \frac{\alpha}{\beta^2}$ e $g(\mu) = log(\mu)$

Este modelo foi obtido através do seguinte código em `R` :

```
mod21 <- with(imp,gam(delay_health ~ sex + age_group + alcohol + inmate +
                      commu_resid + school + house + benef + inc_cat +
                      s(x,y,k=100,bs='tp') + offset(log(n)) , family=Gamma(log)))
```

No contexto da imputação múltipla, utilizamos a função `with` do pacote `mice` para ajustar o modelo GAM a cada um dos conjuntos de dados imputados. A função `with` simplifica a aplicação do modelo a cada imputação, permitindo uma análise mais eficiente e robusta.

De seguida, a função `pool` foi usada para combinar as estimativas dos parâmetros e dos seus erros padrão, considerando a variabilidade entre e dentro das imputações. Este passo é fundamental para obter estimativas finais mais robustas e que representam adequadamente a incerteza associada aos dados omissos. O comando `p <- pool(mod21)` realiza esta combinação. Por fim, utilizando a função `summary(p)` obtivemos um resumo estatístico do nosso modelo final (Tabela 4.9).

Os resultados do modelo final, mostrados na Tabela 4.9, indicam quais os fatores que têm um maior impacto significativo no atraso do diagnóstico nos cuidados de saúde. Este

Variável	Estimativa	Erro Padrão	Estatística	Graus de Liberdade	Valor-P
(Intercept)	4.3695	0.3854	11.3379	12669.2656	1.18×10^{-29}
sex1	0.1373	0.0394	3.4826	10016.9135	4.99×10^{-4}
age_group0	0.0651	0.0874	0.7451	12681.5472	0.4562
age_group2	0.2366	0.0399	5.9236	7497.4856	3.29×10^{-9}
age_group3	0.5266	0.0489	10.7752	12746.0224	5.89×10^{-27}
alcohol1	-0.2749	0.0519	-5.2935	200.0007	3.14×10^{-7}
inmate1	0.4010	0.1611	2.4889	338.9445	0.0133
commu_resid1	-0.2204	0.1016	-2.1706	180.1524	0.0313
school	-0.0818	0.0048	-17.1339	12565.9605	4.54×10^{-65}
house	-0.1418	0.0171	-8.3035	12736.4677	1.11×10^{-16}
benef	-0.0043	0.0013	-3.4566	12773.3906	5.49×10^{-4}
inc_cat1	-0.1582	0.0441	-3.5865	12373.1995	3.36×10^{-4}

TABELA 4.9: Sumário do modelo final GAM com imputação múltipla (IM)

modelo ajuda-nos a entender melhor as variáveis que influenciam este atraso e pode ser útil para desenvolver estratégias de minimização do atraso nos cuidados de saúde no futuro.

Com base na Tabela 4.9 verificamos, em média, um maior tempo de atraso no diagnóstico nos cuidados de saúde em pacientes com as seguintes características: sexo feminino, idade entre os 45 e os 65 anos e superior a 65 anos (última categoria apresenta maiores diferenças no atraso), o facto de ser recluso, sempre quando comparados com as respetivas categorias de referência.

Por outro lado, verificamos, em média, um menor tempo de atraso do diagnóstico nos cuidados de saúde em pacientes: consumidores abusivos de álcool, que vivem numa residência comunitária e que vivem num município com uma taxa de incidência alta, quando comparados com a categoria de referência.

Ainda verificamos também que um aumento da proporção de indivíduos com 3º ciclo de escolaridade, assim como o aumento da proporção de sobrelotação por município e aumento da taxa de beneficiários de rendimento social estão associados a uma diminuição significativa no atraso dos cuidados de saúde.

4.4.4 Eficácia da Imputação Múltipla

A Tabela 4.10 foi obtida através da função `pool(mod21)`, onde estão apresentadas as medidas de informação, como RIV, FMI e RE, que avaliam a eficácia da imputação múltipla.

Variável	m	RIV	FMI	RE
(Intercept)	5	0.0015	0.0016	0.9997
sex1	5	0.0092	0.0093	0.9981
age_group0	5	0.0014	0.0015	0.9997
age_group2	5	0.0149	0.0150	0.9970
age_group3	5	0.0007	0.0009	0.9998
alcohol1	5	0.1630	0.1486	0.9711
inmate1	5	0.1198	0.1122	0.9780
commu_resid1	5	0.1734	0.1571	0.9695
school	5	0.0021	0.0023	0.9995
house	5	0.0009	0.0010	0.9998
benef	5	0.0002	0.0004	0.9999
inc_cat1	5	0.0031	0.0032	0.9994

TABELA 4.10: Medidas de informação para avaliar a eficácia da imputação múltipla

Os valores de RIV e FMI são, em geral, bastante baixos, sugerindo que a variabilidade adicional introduzida pela imputação é mínima para a maioria das variáveis. Isto é um bom indicador da adequação da imputação múltipla para estas variáveis. Adicionalmente, os valores de RE são bastante próximos de 1 para quase todas as variáveis, indicando uma perda de eficiência mínima devido à imputação.

No entanto, variáveis como *alcohol1*, *inmate1* e *commu_resid1* apresentam valores mais elevados de RIV e FMI, sugerindo uma maior incerteza associada à imputação. Estes resultados indicam que as inferências relativas a essas variáveis devem ser interpretadas com mais cautela.

A imputação múltipla aparenta ser adequada para a maioria das variáveis em estudo, no entanto para aumentar a confiabilidade dos resultados poderíamos aumentar o número de imputações.

4.4.5 Comparação modelos

O objetivo desta secção é avaliar o impacto da imputação múltipla nos resultados do modelo, comparando-o com um modelo GAM construído a partir dos dados completos.

A construção do modelo com os dados completos foi realizada de forma análoga ao modelo com imputação múltipla, através do método de seleção de variáveis *backward elimination*.

O modelo final para os dados completos foi obtido com o seguinte código em `R` :

```
mf18<- gam(delay_health ~ sex + age_group + lung_cancer +
```

```
    alcohol + inmate + homeless + commu_resid +
    school + house +  benef + inc_cat +
    s(x,y,k=100,bs='tp') + offset(log(n)),
  data = dados_comp1, family = Gamma(log))
```

Os coeficientes paramétricos deste modelo são apresentados na Tabela 4.11.

Variável	**Estimativa**	**Erro Padrão**	**Estatística**	**Valor-P**
(Intercept)	3.9864	0.4255	9.369	$< 2e-16$
sex1	0.1438	0.0416	3.455	0.0006
age_group0	0.0596	0.0937	0.636	0.5250
age_group2	0.2214	0.0425	5.210	$1.93e-07$
age_group3	0.5669	0.0529	10.710	$< 2e-16$
lung_cancer1	0.4812	0.1995	2.412	0.0159
alcohol1	-0.2657	0.0533	-4.986	$6.27e-07$
inmate1	0.3932	0.1966	2.000	0.0455
homeless1	-0.5640	0.1594	-3.539	0.0004
commu_resid1	-0.3169	0.1117	-2.837	0.0046
school	-0.0789	0.0053	-15.032	$< 2e-16$
house	-0.1225	0.0188	-6.524	$7.15e-11$
benef	-0.0051	0.0014	-3.735	0.0002
inc_cat1	-0.1368	0.0469	-2.914	0.0036

TABELA 4.11: Sumário do Modelo com dados completos

Ao comparar os coeficientes paramétricos dos dois modelos (Tabelas 4.9 e 4.11), notamos diferenças nas variáveis consideradas como estatisticamente significativas. No modelo utilizando os dados completos, variáveis como a indicação de cancro do pulmão e a situação de sem-abrigo tornam-se significativas, realçando o impacto da imputação múltipla nas relações observadas.

No gráfico da Figura 4.11 estão apresentadas as estimativas para as variáveis significativas comuns em ambos os modelos para mais fácil comparação entre os valores das variáveis. A variável residência comunitária é a que apresenta uma maior diferença entre as estimativas dos coeficientes do modelo, algo que faz sentido uma vez que 2.56 % dos seus valores foram imputados.

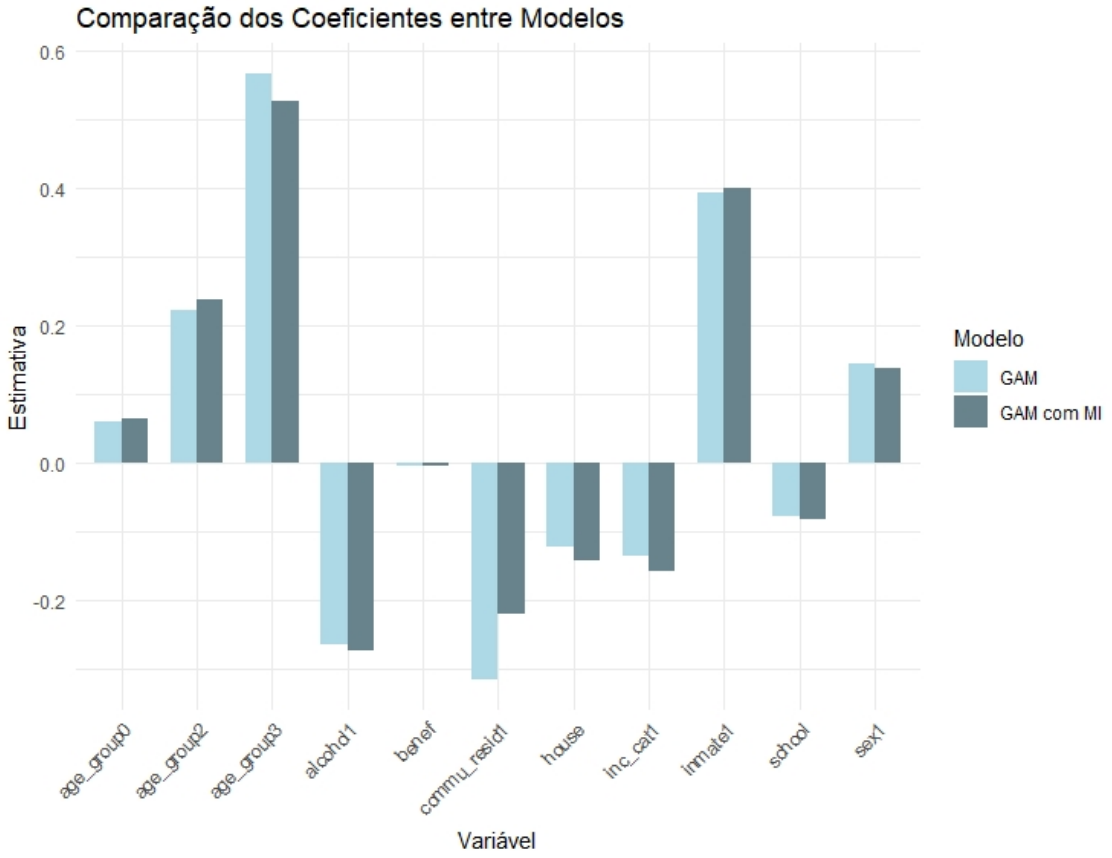


FIGURA 4.11: Comparação entre os coeficientes das variáveis comuns dos modelos GAM com e sem Imputação Múltipla

De seguida, para uma análise mais abrangente, as duas variáveis apenas significativas no modelo sem imputação, nomeadamente a indicação de cancro de pulmão e situação de sem-abrigo, foram adicionadas ao modelo com imputação múltipla.

Assim, o gráfico apresentado na Figura 4.12, abrange todas as variáveis nos dois modelos, destacando a influência e o impacto da imputação múltipla nestas variáveis.

Observam-se diferenças substanciais nas estimativas dos coeficientes destas duas variáveis, revelando o efeito significativo da imputação múltipla nas relações entre as variáveis do modelo.

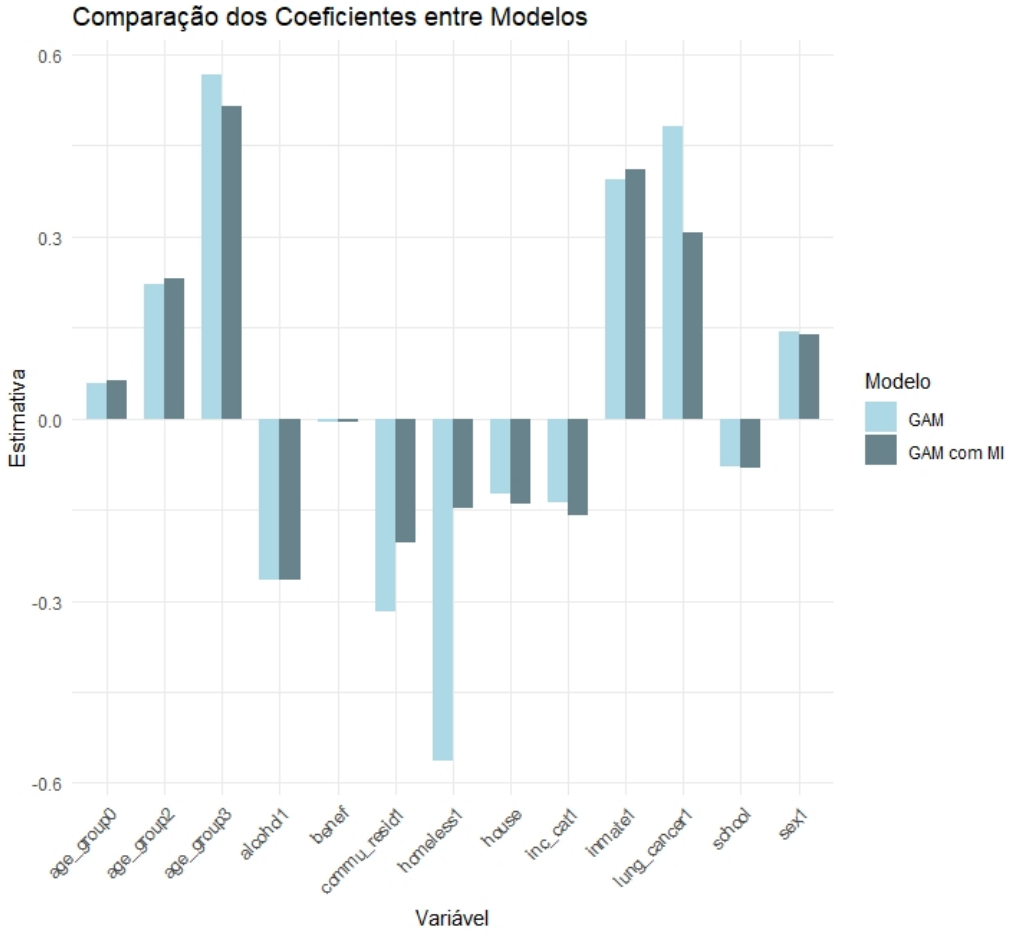


FIGURA 4.12: Comparação das estimativas dos coeficientes para todas as variáveis nos modelos GAM com e sem Imputação Múltipla

A comparação direta entre modelos com e sem imputação não é de todo trivial. Para perceber se efetivamente a imputação múltipla teve um impacto positivo, ou seja, se foi construído um modelo melhor, vamos direcionar o olhar para os intervalos de confiança dos coeficientes paramétricos. As Tabelas seguintes, 4.12 e 4.13, dão conta tanto do erro padrão como dos intervalos de confiança associados aos coeficientes das variáveis comuns aos dois modelos, em ambos os cenários, com e sem imputação.

	GAM		GAM com IM	
Variável	Erro Padrão	Intervalo Confiança	Erro Padrão	Intervalo Confiança
*sex*1	0.0416	[0.0622, 0.2254]	0.0394	[0.0600, 0.2146]
*age_group*0	0.0937	[-0.1241, 0.2433]	0.0874	[-0.1061, 0.2363]
*age_group*2	0.0425	[0.1381, 0.3048]	0.0399	[0.1583, 0.3149]
*age_group*3	0.0529	[0.4632, 0.6707]	0.0489	[0.4308, 0.6224]
*alcohol*1	0.0533	[-0.3701,-0.1612]	0.0519	[-0.3773,-0.1725]
*inmate*1	0.1966	[0.0079, 0.7786]	0.1611	[0.0841, 0.7180]
*commu_resid*1	0.1117	[-0.5359,-0.0980]	0.1016	[-0.4208,-0.0200]
school	0.0053	[-0.0892,-0.0686]	0.0048	[-0.0912,-0.0725]
house	0.0188	[-0.1593,-0.0857]	0.0171	[-0.1753,-0.1083]
benef	0.0014	[-0.0077,-0.0024]	0.0013	[-0.0068,-0.0019]
*inc_cat*1	0.0469	[-0.2288,-0.0448]	0.0441	[-0.2447,-0.0718]

TABELA 4.12: Comparação do Erro padrão e Intervalos de Confiança nos dois modelos

Sabemos que intervalos mais estreitos são indicadores, geralmente, de um modelo mais preciso. Sendo assim, a Tabela 4.13 contém os valores da amplitude dos intervalos de confiança dos dois modelos.

Variável	**GAM**	**GAM com IM**
*sex*1	0.1632	0.1546
*age_group*0	0.3675	0.3425
*age_group*2	0.1566	0.1566
*age_group*3	0.1915	0.1916
*alcohol*1	0.2089	0.2048
*inmate*1	0.7707	0.6339
*commu_resid*1	0.4379	0.4008
school	0.0187	0.0187
house	0.0669	0.0669
benef	0.0053	0.0049
*inc_cat*1	0.1840	0.1730

TABELA 4.13: Comparação da amplitude dos Intervalos de Confiança nos dois modelos

No geral, mesmo que em algumas variáveis a redução seja residual, verifica-se uma diminuição da amplitude dos intervalos de confiança. Tome-se como exemplo a variável inmate, onde a amplitude foi reduzida em aproximadamente 0.14. Sendo assim, estamos perante um ponto a favor da imputação.

Capítulo 5

Conclusão

O principal objetivo deste estudo baseia-se no estudo de modelos de regressão semi-paramétricos, os modelos aditivos generalizados, e analisar e compreender os fatores associados ao tempo de atraso do diagnóstico da tuberculose nos cuidados de sáude. A utilização de modelos aditivos generalizados com termos suaves permitiu capturar relações não lineares e complexas nos dados e o método de imputação múltipla permitiu não ter de excluir observações com valores omissos, de maneira a não sacrificar informação possivelmente valiosa.

Em primeiro lugar, realizamos uma análise descritiva dos dados para perceber como estão distribuídas as variáveis e também verificar as diferenças nos tempos de atraso em alguns fatores. Constatou-se que a maioria dos indivíduos da amostra são do sexo masculino, a faixa etária mais representativa é a de 20 a 44 anos, seguida pela faixa etária entre os 45 a 64 anos. A maioria dos indivíduos é originária de Portugal, em relação à região, a mais representativa é a do Norte, seguida por Lisboa e Vale do Tejo. A maioria dos indivíduos não apresenta condições de saúde como diálise, diabetes, doença hepática, HIV, entre outras. A maioria dos indivíduos não consome álcool ou drogas, não é recluso e nem vive na rua ou numa residência comunitária. A representatividade das categorias da taxa de incidência bruta de cada município na amostra está bastante semelhante. Verificou-se um atraso mediano nos cuidados de saúde superior em pacientes do sexo feminino, com idade superior a 65 anos, originários de Portugal, do Alentejo, empregados, recluso e também em pacientes com a presença de alguma doença, por exemplo diabetes, linfoma, cancro do pulmão, entre outros. Por outro lado, constatou-se um menor atraso mediano nos cuidados de saúde em indivíduos consumidores de álcool e drogas, em indivíduos

que vivem na rua ou numa residência comunitária, e em pacientes que vivem em zonas de alta incidência.

Após a análise exploratória dos dados, foi utilizado o método de imputação múltipla `mice` para preencher os valores omissos e, de seguida, os dados foram interpretados através de um modelo aditivo generalizado, com a função de ligação logarítmica. O modelo final contém essencialmente termos paramétricos, com a exceção da variável espacial que corresponde à componente não paramétrica do modelo. O modelo revela a existência de uma associação positiva da resposta (tempo de atraso nos cuidados de sáude) com o sexo feminino, faixa etária entre 45-64 anos e $\geq$ 65 anos e recluso, em oposição ao sexo masculino, faixa etária entre os 20-44 anos e não ser recluso. Por outro lado, o consumo de álcool, a habitação numa residência comunitária e o domicílio num local com taxa de incidência alta, são variáveis que se revelaram associadas a um atraso inferior, em oposição a não consumidores de álcool, não viver numa residência comunitária e incidência baixa. O aumento da proporção de indivíduos com 3º ciclo de escolaridade, da proporção de sobrelotação e da taxa de beneficiários do rendimento de segurança social estão associados a uma diminuição do tempo de atraso.

Para o modelo aditivo generalizado, com a função de ligação logarítmica, para os dados completos, ou seja sem realizar o método de imputação múltipla, constatou-se algumas diferenças nas variáveis consideradas significativas. Este modelo revela que o aumento do atraso está associado ao sexo feminino, a uma faixa etária entre 45-64 anos e também $\geq$ 65 anos, a pacientes com cancro do pulmão e reclusos, em oposição ao sexo masculino, faixa etária entre 20-44 anos, paciente saudável e não recluso. Por outro lado, o consumo de álcool, a variável indicativa de situação de sem-abrigo, bem como a habitação em residência comunitária e em local com taxa de incidência alta são variáveis que se revelaram associadas a um atraso inferior, em oposição às respetivas categorias de referência. O aumento da proporção de indivíduos com 3º ciclo de escolaridade, da proporção de sobrelotação e da taxa de beneficiários do rendimento de segurança social estão associados a uma diminuição do tempo de atraso.

As conclusões retiradas dos estudos referidos em Roberts et al. [21] e Storla et al. [2], vão ao encontro dos resultados obtidos pelo ajuste de ambos os modelos acima referidos.

A imputação múltipla foi avaliada através de medidas de informação, como RIV, FMI e RE. Estas medidas indicam que a imputação foi eficaz para a maioria das variáveis, com baixa variabilidade adicional introduzida. No entanto, algumas variáveis, como $alcohol$,

inmate e *commu_resid*, mostraram maior incerteza na imputação, sugerindo cautela na interpretação dos resultados relacionados a elas.

O impacto da imputação múltipla foi avaliado comparando o modelo GAM construído com os dados completos e o modelo GAM com imputação múltipla. Observou-se que a imputação influenciou a significância estatística de algumas variáveis. Ao comparar os intervalos de confiança dos coeficientes das variáveis em ambos os modelos, notou-se uma redução geral na amplitude dos intervalos no modelo com imputação, indicando um possível aumento na precisão. Por exemplo, o intervalo de confiança para o coeficiente da variável *inmate* teve a sua amplitude reduzida em cerca de 0.14, destacando o benefício da imputação múltipla.

A imputação múltipla demonstrou ser uma técnica eficaz, proporcionando modelos mais precisos, especialmente quando comparados com modelos construídos a partir de dados completos. No entanto, é essencial considerar a variabilidade e a incerteza associadas à imputação para algumas variáveis específicas.

Estas diferenças realçam a importância de considerar cuidadosamente a abordagem de tratamento de dados omissos em análises estatísticas, especialmente em contextos clínicos e de saúde pública, onde as decisões baseadas nestas análises podem ter implicações significativas. É crucial entender que, enquanto a imputação pode ajudar a maximizar o uso dos dados disponíveis, também pode alterar as relações intrínsecas presentes nos dados originais.

Apesar dos seus benefícios, a imputação múltipla não está isenta de limitações. O método pressupõe que os dados são recolhidos de forma aleatória, o que nem sempre é o caso em estudos reais. Além disso, a qualidade da imputação depende da precisão dos modelos utilizados para estimar os valores omissos. É importante realçar que, a imputação não substitui dados reais e, portanto, haverá sempre alguma incerteza associada às estimativas imputadas.

Num posterior aprofundar da nossa análise, um passo crucial será explorar a eficácia da imputação com um maior número de imputações. Embora tenhamos obtido resultados promissores com o número atual de imputações, aumentar esse mesmo número pode aumentar a precisão e estabilidade das estimativas. Ao fazer isso, esperamos não apenas melhorar a qualidade dos dados imputados, mas também ganhar uma compreensão mais profunda das nuances associadas ao processo de imputação múltipla. Esta abordagem permitir-nos-á avaliar a sensibilidade dos nossos resultados à variação no número

de imputações e otimizar o nosso método para futuras análises.

Apêndice A

Descrição das variáveis

Variável	Variável no R	Descrição
Escolaridade	*school*	Proporção de população com 3º ciclo de escolaridade, por município e sexo (Census 2011)
Casa	*house*	Proporção de sobrelotação, por município (Census 2011)
Beneficiários	*benef*	Beneficiários do rendimento de inclusão social/segurança social por 1000 indivíduos em idade ativa, por município e ano de 2011 a 2017 (INE)
Médicos	*doc_rate*	Médicos em centros de saúde por 10000 indivíduos, por município e ano de 2011 (INE)

TABELA A.1: Descrição variáveis numéricas

Variável	Variável no R	Descrição
Ano	*year*	Ano em que cada paciente foi diagnosticado com tuberculose (entre 2008 e 2017)
Sexo	*sex*	Sexo de cada paciente (Masculino ou Feminino)
Idade	*age_group*	Idade do paciente quando diagnosticado com tuberculose ($< 20; 20-44; 45-64; \geq 65$)

Variável	Variável no R	Descrição
Incidência do país de origem	*country_inc*	País de origem de cada paciente, categorizado pelo grau de incidência de TB (Portugal, País de alta incidência (>20/100 000), País de baixa incidência (<20/100 000))
Distrito	*district*	Distrito de cada paciente
Município	*county*	Município de cada paciente
Código do município	*DICOMUN*	Código de cada município de Portugal
Região	*region*	Região de cada paciente, classificada como Lisboa e Vale do Tejo, Açores, Alentejo, Algarve, Centro, Norte, Madeira
Desemprego	*unemployment*	Indica se o paciente está desempregado há mais de 24 meses
Profissional de saúde	*health_job*	Indica se o paciente é um profissional de saúde
Inconsistências	*problem_health*	Problemas com a data da 1ª consulta, consideramos apenas os casos em que não há inconsistências e os casos em que estão relacionadas com a 1ª consulta e *microscópico* e/ou *cultura*
Incidência	*inc_cat*	Taxa de incidência bruta por 100 000 indivíduos categorizada em baixa (<20/100 000) ou alta (>20/100 000), por município e ano
Diálise	*dialysis*	Indica se o paciente tem insuficiência renal em diálise
Diabetes	*diabetes*	Indica se o paciente tem diabetes
Doença hepática	*liver*	Indica se o paciente tem doença hepática
HIV	*hiv*	Indica se o paciente tem HIV
Doença articular inflamatória	*inflam_joint*	Indica se o paciente tem doença articular inflamatória

Variável	Variável no R	Descrição
Linfoma	*lymphoma*	Indica se o paciente tem linfoma ou doenças mieloproliferativas
Cancro do pulmão	*lung_cancer*	Indica se o paciente tem cancro do pulmão
Outro cancro	*other_cancer*	Indica se o paciente tem outro cancro
Sarcoidose	*sarcoidosis*	Indica se o paciente tem sarcoidose
DPOC	*COPD*	Indica se o paciente tem doença pulmonar obstrutiva crônica
Silicose	*silicosis*	Indica se o paciente tem silicose
Intersticial	*interstitial*	Indica se o paciente tem outra doença intersticial pulmonar
Outra doença	*other_disease*	Indica se o paciente tem outra doença
Álcool	*alcohol*	Indicador de consumo abusivo de álcool
Drogas	*drugs*	Indicador de consumo abusivo de drogas
Recluso	*inmate*	Indica se o paciente é recluso
Sem abrigo	*homeless*	Indica se o paciente é sem abrigo
Residência comunitária	*commu_resid*	Indica se o paciente vive numa residência comunitária

TABELA A.2: Descrição variáveis categóricas

Apêndice B

Descrição numérica das variáveis categóricas

Variável	N = 12 882	Tempo de Atraso mediano nos C.S. (dias)	Valores omissos
Sexo			0.00%
0 - Masculino	9067 (70.39%)	8 (1-31)	
1 - Feminino	3815 (29.61%)	11 (2-40)	
Idade			0.08%
0 - < 20	533 (4.14%)	8 (2-28)	
1 - $20 - 44$	5809 (45.13%)	7 (1-26)	
2 - $45 - 64$	4288 (33.31%)	8 (1-33)	
3 - ≥ 65	2242 (17.42%)	21 (4-54)	
Incidência do país de origem			0.06%
0 - Portugal	10852 (84.29%)	9 (1-34)	
1 - Alta	899 (6.98%)	8 (1-29)	
2 - Baixa	1123 (8.72%)	7 (1-27)	
Região			0.00%
0 - Lisboa e Vale do Tejo	5085 (39.47%)	9 (1-35)	
1 - Alentejo	482 (3.74%)	10.5 (1-47)	
2 - Algarve	527 (4.09%)	9 (2-31)	

Variável	**N = 12 882**	**Tempo de Atraso mediano nos C.S. (dias)**	**Valores omissos**
3 - Centro	1090 (8.46%)	7 (1-33)	
4 - Norte	5698 (44.23%)	8 (1-32)	
Desemprego			0.00%
0 - Não	10787 (83.74%)	9 (1-35)	
1 - Sim	2095 (16.26%)	6 (1-22)	
Profissional de saúde			0.00%
0 - Não	12359 (95.94%)	8 (1-33.50)	
1 - Sim	523 (4.06%)	8 (1-31.50)	
Diálise			0.00%
0 - Não	12773 (99.15%)	8 (1-33)	
1 - Sim	63 (0.85%)	21 (4-46)	
Diabetes			0.00%
0 - Não	12031 (93.39%)	8 (1-33)	
1 - Sim	851 (6.61%)	11 (2-41)	
Doença hepática			0.00%
0 - Não	12248 (95.08%)	8 (1-33)	
1 - Sim	634 (4.92%)	9 (1-35)	
HIV			8.33%
0 - Não	10411 (88.30%)	8 (1-34)	
1 - Sim	1379 (11.70%)	8 (1-27)	
Doença articular inflamatória			1.78%
0 - Não	12546 (99.19%)	8 (1-33)	
1 - Sim	103 (0.81%)	16 (2-51.50)	
Linfoma			0.00%
0 - Não	12819 (99.51%)	8 (1-33)	
1 - Sim	63 (0.49%)	29 (6.5-54.5)	

Variável	N = 12 882	Tempo de Atraso mediano nos C.S. (dias)	Valores omissos
Cancro do pulmão			0.00%
0 - Não	12756 (99.02%)	8 (1-33)	
1 - Sim	126 (0.98%)	32 (7.25-59.75)	
Outro cancro			0.00%
0 - Não	12483 (96.90%)	8 (1-33)	
1 - Sim	399 (3.10%)	20 (4-48)	
Sarcoidose			0.00%
0 - Não	12863 (99.85%)	8 (1-33)	
1 - Sim	19 (0.15%)	28 (7.5-61.5)	
COPD			0.00%
0 - Não	12377 (96.08%)	8 (1-33)	
1 - Sim	505 (3.92%)	13 (2-51)	
Silicose			0.00%
0 - Não	12662 (98.29%)	8 (1-33)	
1 - Sim	220 (1.71%)	11 (3.75-38.25)	
Intersticial			0.00%
0 - Não	12825 (99.56%)	8 (1-33)	
1 - Sim	57 (0.44%)	20 (2-52)	
Outra doença			0.00%
0 - Não	11046 (85.75%)	8 (1-33)	
1 - Sim	1836 (14.25%)	11 (2-38)	
Álcool			4.41%
0 - Não	10299 (83.70%)	10 (1-36)	
1 - Sim	2005 (16.30%)	5 (1-22)	
Drogas			3.64%
0 - Não	11064 (89.19%)	9 (1-35)	
1 - Sim	1341 (10.81%)	6 (1-24)	

Variável	**N = 12 882**	**Tempo de Atraso mediano nos C.S. (dias)**	**Valores omissos**
Recluso			2.21%
0 - Não	12439 (98.78%)	9 (1-34)	
1 - Sim	153 (1.22%)	15 (2-38)	
Sem abrigo			2.25%
0 - Não	12350 (98.12%)	9 (1-34)	
1 - Sim	237 (1.88%)	4 (1-18)	
Residência comunitária			2.56%
0 - Não	12139 (96.76%)	9 (1-34)	
1 - Sim	407 (3.24%)	6 (1-23)	
Incidência			0.00%
0 - Baixa	6678 (51.84%)	10 (2-37)	
1 - Alta	6204 (48.16%)	8 (1-30)	

TABELA B.1: Descrição numérica - variáveis categóricas

Apêndice C

Correlação

Variável	Valor t	valor-p	Correlação	Intervalo de Confiança (95%)
sex	1.6195	0.1054	0.0143	[-0.0030 , 0.0315]
unemployment	6.2427	4.435×10^{-10}	0.0549	[0.0377 , 0.0721]
health_job	-4.5945	4.379×10^{-6}	-0.0405	[-0.0577 , -0.0232]
dialysis	-0.9020	0.3671	-0.0079	[-0.0252 , 0.0093]
lymphoma	1.9007	0.0574	0.0167	[-0.0005 , 0.0340]
lung_cancer	1.7719	0.0764	0.0156	[-0.0017 , 0.0329]
other_cancer	2.1551	0.0312	0.0190	[0.0017 , 0.0362]
diabetes	0.5998	0.5486	0.0053	[-0.0120 , 0.0226]
liver	4.4754	7.691×10^{-6}	0.0394	[0.0222 , 0.0566]
sarcoidosis	0.6899	0.4902	0.0061	[-0.0112 , 0.0233]
inflam_joint	3.5788	0.0003	0.0318	[0.0144 , 0.0492]
COPD	4.6038	4.189×10^{-6}	0.0405	[0.0233 , 0.0578]
silicosis	-1.5427	0.1229	-0.0136	[-0.0309 , 0.0037]
interstitial	1.5003	0.1336	0.0132	[-0.0041 , 0.0305]
HIV	0.5253	0.5994	0.0048	[-0.0132 , 0.0229]
other_disease	11.224	$< 2.2 \times 10^{-16}$	0.0984	[0.0813 , 0.1155]
alcohol	2.4142	0.0158	0.0218	[0.0041 , 0.0394]
drugs	2.6719	0.0076	0.0240	[0.0064 , 0.0416]
inmate	0.8604	0.3896	0.0077	[-0.0098 , 0.0251]
homeless	3.3211	0.0009	0.0296	[0.0121 , 0.0470]
commu_resid	2.3338	0.0196	0.0208	[0.0033 , 0.0383]
problem_health	-1.0199	0.3078	-0.0090	[-0.0263 , 0.0083]
inc_cat	16.721	$< 2.2 \times 10^{-16}$	0.1458	[0.1288 , 0.1626]

TABELA C.1: Resumo dos resultados de correlação.

Realizamos uma ANOVA para avaliar como *doc_rate* varia entre diferentes grupos de variáveis categóricas com mais de dois níveis. Os resultados, apresentados na Tabela C.2, mostram que o *doc_rate* varia significativamente entre diferentes grupos etários, incidência do país de origem e região de cada paciente.

Variável	**Graus de liberdade**	**Soma dos Quadrados**	**Quadrado Médio**	**Valor F**	$Pr(>F)$
age_group	3	55	18.321	7.011	0.000104
country_inc	2	257	128.3	49.38	$<$ 2e-16
region	4	3486	871.6	371.3	$<$ 2e-16

TABELA C.2: Resultados da ANOVA para variáveis categóricas.

Bibliografia

[1] W. H. Organization *et al.*, "Global tuberculosıs report 2022," *World Health Organization, Geneva, Switzerland*, 2022. [Cited on page 1.]

[2] D. G. Storla, S. Yimer, and G. A. Bjune, "A systematic review of delay in the diagnosis and treatment of tuberculosis," *BMC public health*, vol. 8, no. 1, pp. 1–9, 2008. [Cited on pages 1 and 54.]

[3] J. A. Santos, A. Leite, P. Soares, R. Duarte, and C. Nunes, "Delayed diagnosis of active pulmonary tuberculosis-potential risk factors for patient and healthcare delays in portugal," *BMC Public Health*, vol. 21, no. 1, pp. 1–13, 2021. [Cited on page 1.]

[4] R. K. Mahato, W. Laohasiriwong, K. Vaeteewootacharn, R. Koju, and R. Bhattarai, "Major delays in the diagnosis and management of tuberculosis patients in nepal," *Journal of clinical and diagnostic research: JCDR*, vol. 9, no. 10, p. LC05, 2015. [Cited on page 1.]

[5] S. O. Olarewaju, O. A. Alawode, O. T. Adegbosin, A. B. Olaniyan, and S. C. Adeyemo, "Factors that influence diagnostic delay among pulmonary tuberculosis patients in osogbo, nigeria," *Journal of the Pan African Thoracic Society*, vol. 4, no. 1, pp. 22–30, 2023. [Cited on page 1.]

[6] M. G. Farah, J. H. Rygh, T. W. Steen, R. Selmer, E. Heldal, and G. Bjune, "Patient and health care system delays in the start of tuberculosis treatment in norway," *BMC infectious diseases*, vol. 6, pp. 1–7, 2006. [Cited on pages 1 and 2.]

[7] P. Tattevin, D. Che, P. Fraisse, C. Gatey, C. Guichard, D. Antoine, M. Paty, and E. Bouvet, "Factors associated with patient and health care system delay in the diagnosis of tuberculosis in france," *The international journal of tuberculosis and lung disease*, vol. 16, no. 4, pp. 510–515, 2012. [Cited on pages 1 and 2.]

[8] M. Loutet, C. Sinclair, N. Whitehead, C. Cosgrove, M. Lalor, and H. Thomas, "Delay from symptom onset to treatment start among tuberculosis patients in england, 2012–2015," *Epidemiology & Infection*, vol. 146, no. 12, pp. 1511–1518, 2018. [Cited on pages 1 and 2.]

[9] A. M. Peri, D. P. Bernasconi, N. Galizzi, A. Matteelli, L. Codecasa, V. Giorgio, A. Di Biagio, F. Franzetti, A. Cingolani, A. Gori *et al.*, "Determinants of patient and health care services delays for tuberculosis diagnosis in italy: a cross-sectional observational study," *BMC infectious diseases*, vol. 18, pp. 1–11, 2018. [Cited on pages 1 and 2.]

[10] A. Seminario, L. Anibarro, J. Sabriá, M. M. García-Clemente, A. Sánchez-Montalván, J. F. Medina, I. Mir, A. Penas, J. A. Caminero, G. J. Pérez *et al.*, "Study of the diagnostic delay of tuberculosis in spain," *Archivos de bronconeumologia*, vol. 57, no. 6, pp. 440–442, 2021. [Cited on page 2.]

[11] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*. CRC press, 1990, vol. 43. [Cited on pages 2 and 13.]

[12] D. B. Rubin, "Multiple imputation for survey nonresponse," 1987. [Cited on pages 2, 7, 9, and 10.]

[13] K. Kleinke, J. Reinecke, D. Salfrán, and M. Spiess, *Applied multiple imputation*. Springer, 2020. [Cited on page 5.]

[14] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, pp. 1–17, 2013. [Cited on page 7.]

[15] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, vol. 45, pp. 1–67, 2011. [Cited on pages ix and 7.]

[16] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?" *International journal of methods in psychiatric research*, vol. 20, no. 1, pp. 40–49, 2011. [Cited on page 8.]

[17] J. W. Graham, A. E. Olchowski, and T. D. Gilreath, "How many imputations are really needed? some practical clarifications of multiple imputation theory," *Prevention science*, vol. 8, pp. 206–213, 2007. [Cited on page 9.]

[18] M. Heymans and I. Eekhout, "Applied missing data analysis with spss and (r) studio," *Heymans and Eekhout: Amsterdam, The Netherlands: 20Available online: https://bookdown. org/mwheymans/bookmi/[accessed 23 May 2020]*, 2019. [Cited on page 10.]

[19] S. N. Wood, *Generalized additive models: an introduction with R*. chapman and hall/-CRC, 2006. [Cited on pages ix, 13, 15, 16, 17, 18, 19, 21, 26, and 27.]

[20] C. Gu and C. Gu, *Smoothing spline ANOVA models*. Springer, 2013, vol. 297. [Cited on page 15.]

[21] D. J. Roberts, T. Mannes, N. Q. Verlander, and C. Anderson, "Factors associated with delay in treatment initiation for pulmonary tuberculosis," *ERJ open research*, vol. 6, no. 1, 2020. [Cited on page 54.]