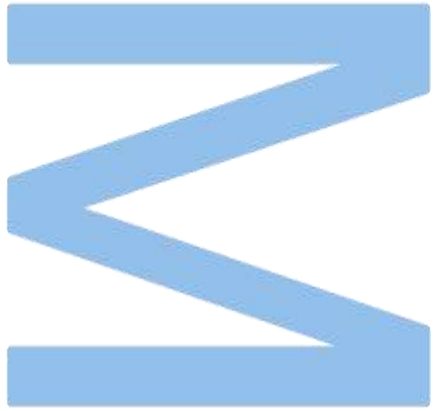
Addressing Missing Data in the Development of a Risk Prediction Model for Childhood Obesity



Mafalda Oliveira
Mestrado em Estatística Computacional e Análise de Dados
Departamento de Matemática da Faculdade de Ciências da Universidade do Porto
2023

Supervisor
Ana Rita Pires Gaio, Professora Auxiliar,
Faculdade de Ciências da Universidade do Porto


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Sworn Statement

I, Mafalda Inês Teixeira Oliveira, enrolled in the Master Degree in Computational Statistics and Data Analysis at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this internship report reflects perspectives, research work and my own interpretations at the time of its submission. By submitting this internship report, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution. I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This internship report does not include any content whose reproduction is protected by copyright laws. I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Mafalda Inês Teixeira Oliveira June 30, 2023

Acknowledgements

First of all, I have to thank my supervisor, Professor Rita Gaio, for her support, interest, patience, and constant availability. Thank you for continually inspiring me to produce the best possible work and for consistently pushing me to exceed my limits. I learned a lot from you in this journey, not only academically but also personally. Having Professor Rita as a supervisor made the whole process of writing this report more interesting and it was a pleasure to work with her. I am certain that I could not have chosen a better supervisor.

Secondly, I would also like to thank my co-supervisor, Professor Susana Santos for her constant support, understanding, and constant availability throughout this internship. Susana's guidance was fundamental for my integration in the internship and I am very grateful for having had the opportunity to learn from you and to learn more about epidemiology.

I would also like to thank all my friends for their support and for always believing in me. Special thanks to Daniel for all his patience and motivation during the whole process, and to Francisca not only for her company during the internship but also for her patience and affection.

Special thanks also to my colleague Rui Miranda for his motivation and availability to help, especially in the final stage of writing the report.

Finally, I also want to thank my family for their unconditional support and motivation. Their unconditional love and encouragement have been vital throughout this journey.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Matemática

MSc. Computational Statistics and Data Analysis

Addressing Missing Data in the Development of a Risk Prediction Model for Childhood Obesity

by Mafalda OLIVEIRA

Obesity is a complex and multifactorial disease that has emerged as a significant public health concern. One meaningful way to address the obesity epidemic is to identify at-risk individuals in advance. This enables early intervention and the implementation of preventive measures. Previous studies have explored the simultaneous association of social factors, lifestyle behaviors, and anthropometric measures with childhood obesity. Research that contemplates a more holistic approach studying the concept of exposome obesity in a dynamic model, is still lacking.

The aim of the present report is to systematically assess the associations between a wide array of exposures measured prenatally and during childhood, and obesity, in order to build a dynamic early-life exposome prediction model for the early identification of individuals susceptible to obesity at 13 years old. The study used data from the Generation XXI cohort, including 4246 mother-child pairs. The children were evaluated at 2, 4, 7, and 10 years old providing a total of 55 collected exposures, including sociodemographic, anthropometric, and lifestyle exposures, related not only to the child but also to the mother's pregnancy stage. Due to the presence of missing values in the dataset, a careful analysis of the available imputation methods was performed before fitting the model. Multiple Imputation with Chained Equations was the selected method to impute the missing values. The dynamic model aimed to predict obesity at age 13 through static and longitudinal predictors measured at 2, 4, 7, and 10 years old. Due to the cross-sectional nature of the response variable, traditional longitudinal models were not applicable. Therefore, three different approaches were explored to address this challenge. Firstly, four separate

logistic regression models were fitted, one for each time point, to investigate the associations between exposures at that time point and the outcome. Secondly, a dynamic model incorporating all exposures, collected at all available time points, was applied using a penalized regression method with Elastic Net (ENET). The models' performances were evaluated by the Area Under the ROC Curve (AUC), Specificity, Sensitivity, Prediction Error, Positive Predictive Ability, and Negative Predictive Ability. Lastly, a finite mixture of penalized regressions with Least Absolute Shrinkage and Selection Operator (LASSO), was fitted and the results were evaluated using the mean squared error, mean absolute error, and mean absolute percentage error.

The results indicated that the Child's Body Mass Index measured at each follow-up was systematically the most influential predictor for obesity at age 13. Additionally, several sociodemographic exposures related to the mother exhibited significant associations. The results from the four static models suggest that the inclusion of the most recently collected predictors improves the performance measures at most. The dynamic model with ENET regression presented satisfactory performance results, with values very similar to those obtained for the 10 years old model (17% of prediction error). The mixture of regressions outperformed the previous models, with a mean absolute percentage error of only 5%.

The built models proved to be efficient in predicting childhood obesity at 13 years old. Early prevention should focus on the mother's body mass index before pregnancy and the child's body mass index across childhood.

Key-words: childhood obesity, finite mixture models, logistic regression, missing values, multiple imputation, penalized regression

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Matemática

Mestrado em Estatística Computacional e Análise de dados

Modelo dinâmico para a previsão de obesidade infantil considerando o expossoma

por Mafalda OLIVEIRA

A obesidade é uma doença complexa e multifatorial que emergiu como um importante problema de saúde pública. Uma forma significativa de abordar a epidemia de obesidade é identificar antecipadamente os indivíduos em risco. Isto permite uma intervenção precoce e a implementação de medidas preventivas. Estudos anteriores exploraram a associação simultânea de fatores sociais, comportamentos de estilo de vida e medidas antropométricas com a obesidade infantil. No entanto, faltam ainda investigações que contemplem uma abordagem mais holística, que estude o conceito do expossoma na obesidade num modelo dinâmico.

O objetivo deste relatório é avaliar sistematicamente as associações entre um vasto leque de exposições medidas no período pré-natal e durante a infância com a obesidade, a fim de desenvolver um modelo dinâmico de previsão usando a abordagem do expossoma para a identificação precoce de indivíduos suscetíveis a ter obesidade aos 13 anos de idade. O estudo utilizou dados da coorte Geração XXI, incluindo 4246 pares mãe-filho. As crianças foram avaliadas aos 2, 4, 7 e 10 anos de idade, fornecendo um total de 55 exposições recolhidas, incluindo exposições sociodemográficas, antropométricas e de estilo de vida, relacionadas não só com a criança mas também com a gravidez da mãe. Devido à presença de valores omissos no conjunto de dados, foi efetuada uma análise cuidadosa dos métodos de imputação disponíveis antes de ajustar o modelo. A imputação múltipla por equações encadeadas foi o método selecionado para imputar os valores omissos. O modelo dinâmico tinha como objetivo prever a obesidade aos 13 anos de idade através de preditores estáticos e longitudinais medidos aos 2, 4, 7 e 10 anos de

idade. Devido à natureza transversal da variável resposta, os modelos longitudinais tradicionais não foram aplicados. Por conseguinte, foram exploradas três abordagens diferentes para responder a este desafio. Em primeiro lugar, foram ajustados quatro modelos de regressão logística separados, um para cada *time-point*, para investigar as associações entre as exposições nesse momento e a variável resposta. Em segundo lugar, foi aplicado um modelo dinâmico que incorporava todas as exposições, recolhidas em todos os pontos temporais disponíveis, utilizando um método de regressão penalizada com *Elastic Net* (ENET). Os desempenhos dos modelos foram avaliados pela área sob a curva ROC (AUC), especificidade, sensibilidade, erro de previsão, capacidade de previsão positiva e capacidade de previsão negativa. Por último, foi ajustada uma mistura finita de regressões penalizadas com *Least Absolute Shrinkage and Selection Operator* (LASSO) e os resultados foram avaliados através do erro quadrático médio, do erro absoluto médio e do erro percentual absoluto médio.

Os resultados indicaram que o Índice de Massa Corporal da criança medido em cada acompanhamento foi sistematicamente a variável mais influente para a obesidade aos 13 anos. Além disso, várias exposições sociodemográficas relacionadas com a mãe apresentaram associações significativas. Os resultados dos quatro modelos estáticos sugerem que a inclusão dos preditores recolhidos mais recentemente melhora as medidas de desempenho. O modelo dinâmico com regressão ENET apresentou resultados de desempenho satisfatórios, com valores muito semelhantes aos obtidos para o modelo dos 10 anos (17% de erro de predição). A mistura de regressões superou os modelos anteriores, com um erro percentual absoluto médio de apenas 5%.

Os modelos construídos revelaram-se eficientes na previsão da obesidade infantil aos 13 anos de idade. A prevenção precoce deve centrar-se no índice de massa corporal da mãe antes da gravidez e no índice de massa corporal da criança ao longo da infância.

Palavras-chave: imputação múltipla, modelo de misturas de regressão, obesidade infantil, regressão com penalização, regressão logistica, valores omissos

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Internship at ISPUP

This report resulted from an internship at the *Instituto de Saúde Pública Da Universidade do Porto* (ISPUP). During this internship, under the guidance of my co-advisor, Professor Susana Santos, I had the privilege of working on a project focused on childhood obesity. The main objective of the internship was to develop a dynamic predictive model for childhood obesity using an exposome approach. Throughout the internship, Professor Susana provided valuable insights and motivation that shaped the additional objectives of this project.

1.2 Introduction

Obesity is a complex multifactorial disease that has become one of the most alarming public health diseases [1]. According to the World Health Organization (WHO), worldwide overweight and obesity prevalence in children and adolescents has increased from 4% to 18% between 1975 and 2016 [2]. These rising values translate into a significant health and economic burden. Being overweight and obese in childhood has been associated with a higher risk of developing type 2 diabetes, cardiovascular diseases, certain cancers, and adverse mental health outcomes in childhood and later in life [3] [4] [5] [6]. Moreover, obesity tracks over time: 55% of children will continue to have obesity in adolescence, and 80% of adolescents with obesity will maintain it in their adult life [7]. It is now estimated that in the coming 30 years, diseases caused by obesity will account for 8.4% of

the total healthcare spending in countries belonging to the Organization for Economic Co-operation and Development (OECD) [8].

One important way to address the obesity epidemic is by identifying in advance individuals who are at risk of becoming obese. This allows for early intervention and the implementation of preventive measures. In fact, tackling obesity in adolescents may help prevent the disease from continuing into adulthood. Previous literature has identified factors such as genetic background, maternal and familiar behaviors, characteristics, lifestyle behaviors (including sedentary, sleep, and dietary habits), as well as other environmental and social influences as important determinants of obesity [3] [9] [10]. However, the existing studies were limited as they only considered a few exposures and explored them at a single time point. To fill this gap, Dr. Christopher Wild introduced the concept of exposome, in 2005 [11]. It was described as "the totality of human environmental exposures from conception onwards", recognizing that individuals are exposed simultaneously to a multitude of different factors across several life stages and thus taking a holistic approach to the discovery of etiological factors [12]. Wild argued that it is necessary to consider the wide range of environmental exposures that an individual experiences throughout his/her lifetime, and how these exposures interact with genetics, in order to achieve a better understanding of health and diseases. The exposome consists of three domains: a general external domain that includes exposures such as climate, urban/rural environment, and socio-economic and psychological factors; a specific external domain that consists of environmental pollutants, radiation, chemical agents, and lifestyle factors (assessed at the individual level); and an internal domain with factors about internal body processes (inflammation, metabolism, hormones, oxidative stress) [13]. Figure 1.1 (adapted from [14]) shows the exposome domains.

Even in its partial forms, the exposome provides a useful framework to systematically evaluate many associations and may be used to avoid problems of selective reporting, publication bias, and confounding by co-exposures, ingrained in the traditional one-by-one reporting of associations [13]. In particular, exposome-obesity analyses may help both in the discovery of novel risk factors and in setting priorities for prevention. Up to now, the study by Vrijheid et al. ([15]) is the only published research considering exposome obesity in childhood. However, this study is cross-sectional, dismissing the time-variability component of the exposures throughout time. Besides the identification of the potentially modifiable risk factors, the prevention of obesity also requires the timely identification of

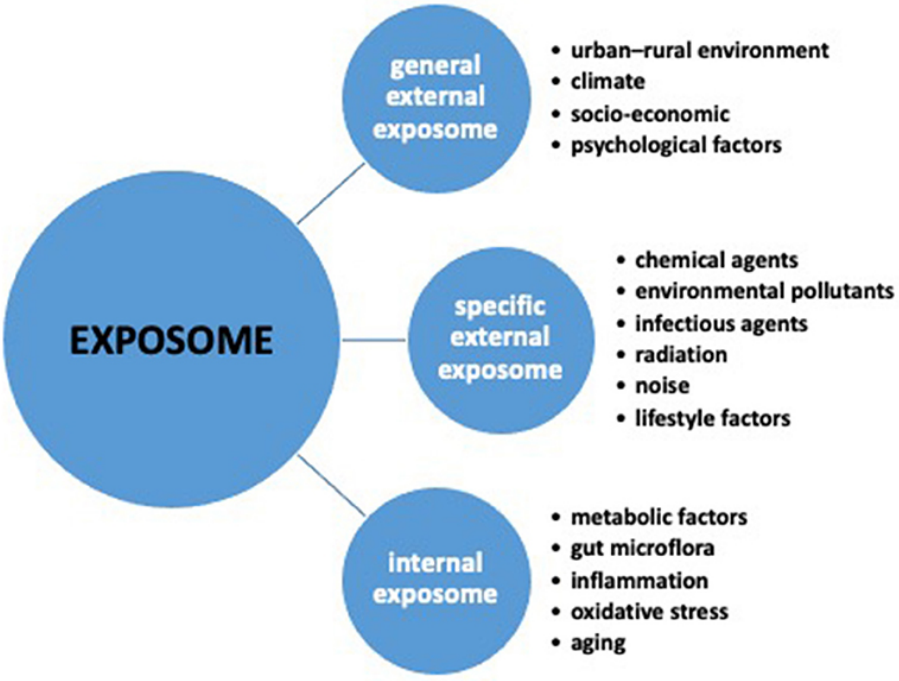


FIGURE 1.1: Exposome domains.

individuals at increased risk. To the best of my knowledge, existing prediction models use a reductionistic approach by focusing on a small subset of risk factors, and/or poorly incorporating their dynamic nature [16] [17] [18][19] [20]. Consequently, factors forthcoming from the exposome may contribute to developing clinical dynamic prediction models to identify individuals at risk of obesity from the earliest phase of life. Hence, it is necessary to contemplate the exposome obesity in a dynamic model to study different time points of the patient.

This internship report aims to apply an exposome approach to systematically assess the associations between a wide array of exposures measured prenatally and during childhood at different time points. The analysis will be conducted on the study population from the Generation XXI (G21) cohort. This is a population-based birth cohort in the Porto Metropolitan Area that, between April 2005 and August 2006, recruited 8647 newborns and their mothers from the five level-III public maternities in this area [21]. More details regarding the dataset used in the analysis will be presented later.

Our exposome approach brought along some challenges from the statistical point of view. Firstly, we refer to the presence of missing values, as a common challenge in data analysis, particularly in the domain of public health. Missing values concern a lack of information in the dataset which is a result of a missing record or an incorrect record for

a particular variable in a specific observation. The presence of missing values has a significant impact on the conclusions that can be drawn from a research study since it can lead to biased or incorrect results [22]. Several methods such as maximum likelihood, weighted estimating equations, single imputation, or multiple imputation were developed to deal with missing data [23] [24] [22]. While several papers compare imputation methods through simulation studies ([25] [26] [27]), the use of Multiple Imputation with Chained Equation (MICE) is particularly noteworthy due to its flexibility, accuracy, and ability to handle different missing data patterns [28]. By generating multiple imputations and accounting for the uncertainty inherent to the imputation process, MICE provides a robust solution for dealing with missing data.

Variable selection is an important methodology to consider in our epidemiologic framework since the exposome approach includes a large number of variables. Many variable selection techniques such as subset selection, regularization methods, and information criteria will be approached throughout this report [29] [30].

A comprehensive understanding of the long-term factors that contribute to obesity in childhood requires the inclusion of longitudinal variables measured at different time points (e.g. at pregnancy, infancy, 4, 7, and 10 years old). However, a problem arises when the response variable is cross-sectional, as we have in our setting. Previous studies describe some methods to incorporate longitudinal predictors with a cross-sectional outcome [31] [32] [33]. We will describe them in later chapters. We will also build a dynamic early-life exposome prediction model for the early identification of individuals susceptible to obesity at 13 years old. The analysis will be divided into two stages. Firstly, we will consider static models focusing on the association between the exposures collected during pregnancy/infancy, 4, 7, and 10 years old, and the outcome variable (obesity at age 13), separately for each time point. Secondly, a dynamic model will predict obesity at age 13 based on static and longitudinal predictors. The static models will be compared with the dynamic model. Furthermore, the applicability of the models for clinical purposes will be discussed.

All statistical analyses were performed in R (version 4.3.0) and the significance level was set at 0.05.

1.3 Structure

This report is divided into the following chapters:

Chapter 1 - Introduction: This chapter introduces the research topic, provides background information, and outlines the objectives of this report.

Chapter 2 - Missing Values: This chapter presents the topic of missing values. I discuss the mechanisms of missing data and present various methods for handling missing values. The focus will be on multiple imputation techniques, particularly on the Multiple Imputation with Chained Equations (MICE) method. Some R packages to handle missing values are presented.

Chapter 3 - Variable Selection: This chapter introduces different variable selection methods, including feature selection (supervised) methods: filter models, wrapper models, and embedded models. Additionally, it addresses how variable selection methods can be applied after multiple imputation has been performed.

Chapter 4 - Methods for Incorporating Longitudinal Predictors with a Static Outcome: This chapter focuses on the challenge of incorporating longitudinal predictors in a regression model with a static response model. I will present some simple approaches and describe finite mixtures of regression models.

Chapter 5 - Dataset: This chapter provides a comprehensive description of the dataset used in the study. I will outline the data collection procedure, and describe the variables included. A descriptive analysis is then performed.

Chapter 6 - Results - Missing Values: This chapter presents and discusses the results obtained from the imputation procedure. Firstly, a description of the missing values and their patterns in the dataset is performed. Secondly, analyses are conducted in order to compare the results from different imputation methods.

Chapter 7 - Results - Models: In this chapter, the results from the four static models and the dynamic model are presented. A comparison of the model's performances is also addressed.

Chapter 8 - Conclusions: This chapter includes a summary and a discussion of the main findings and presents lines of future research.

Chapter 2

Missing Values

Dealing with missing values is one of the most inconvenient problems in data analysis. Several research areas such as statistics, marketing, economics, medical and health data, surveys, and self-reported data have been studying and handling datasets with missing values. This terminology refers to the lack of information in the dataset which is a result of a missing record or an incorrect record for a particular variable in a specific observation. The presence of missing values has a significant impact on the conclusions that can be drawn from a research study since it can lead to biased or incorrect results. Additionally, the pattern of the missing observations (random or systematic) has serious implications on how to appropriately handle missing data since most missing data techniques depend on the missingness mechanisms [22].

The missing values mechanisms can be classified into three distinct categories: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) [34][22]. Let $Y = (y_{ij})$ be an $(n \times K)$ matrix representing a dataset with n observations and K variables, where i represents the i^{th} observation, j the j^{th} variable and denote the i^{th} row by y_i. Let $M = (m_{ij})$ be an $(n \times K)$ matrix where $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed, and let $O_i = \{j : m_{ij} = 0\}$ be the set of indexes of observed variables for subject i, and $R_i = \{j : m_{ij} = 1\}$ the set of indexes of missing variables for subject i. It is possible to write $Y_i = (Y_{iO_i}, Y_{iR_i})$. For simplicity, let us assume that the rows (y_i, m_i) are independent and identically distributed over i. Moreover, let $f_{M|Y}(m_i|y_i, \theta)$ be the conditional distribution of m_i given y_i, with θ representing the vector of unknown parameters. MCAR occurs whenever the probability of a value being missing is the same for all cases and is not related to the missing value itself or any other variable.

This means that the missing value is independent of both observed and unobserved data. So, for all i and any distinct row vectors y_i, y_i^* in the sample space of Y [22]:

$$f_{M|Y}(m_i|y_i,\theta) = f_{M|Y}(m_i|y_i^*,\theta) \tag{2.1}$$

Equivalently, if we denote $f_{M|Y}(m_i|y_i,\theta)$ by $P(M_i|Y_i)$, $f_Y(y_i)$ by $P(Y_i)$, and $f_M(m_i)$ by $P(M_i)$, equation (2.1) becomes equation (2.2):

$$P(M_i|Y_i) = P(M_i) \tag{2.2}$$

We now prove that the results from a complete case analysis (which means excluding all observations with missing values) under the MCAR assumption are unbiased [24]:

$$P(Y_i|M_i=0) = \frac{P(M_i=0|Y_i)P(Y_i)}{P(M_i=0)} = \frac{P(M_i=0)P(Y_i)}{P(M_i=0)} = P(Y_i)$$

that is, the distribution of the complete cases (Y_i when $M_i = 0$) is equal to the distribution of the entire row (Y_i). Hence, MCAR is the preferred missing mechanism for statistical analyses.

The MAR mechanism occurs whenever the probability of a value being missing depends on other variables in the dataset but is independent of unobserved data. So let $y_{(1)i}$ represent the values of the variables for which observation i has a missing value, and $y_{(0)i}$ the remaining values in the dataset. The dataset is MAR if, for all i and any distinct values $(y_{(1)i}, y_{(1)i}^*)$ of the missing components in the sample space of $y_{(1)i}$ [22]:

$$f_{M|Y}(m_i|y_{(0)i},y_{(1)i},\theta) = f_{M|Y}(m_i|y_{(0)i},y_{(1)i}^*,\theta)$$

or, equivalently:

$$P(M_i|Y_i) = P(M_i|Y_{(0)i})$$

In datasets with a MAR mechanism, a complete case analysis should not be performed since it can lead to biased results [22]. In fact [24],

$$P(Y_i|M_i=0) = \frac{P(M_i=0|Y_i)P(Y_i)}{P(M_i=0)} = \frac{P(M_i=0|Y_{(0)i})P(Y_i)}{P(M_i=0)}$$

which is different from $P(Y_i)$ since $P(M_i=0|Y_{(0)i}) \neq P(M_i=0)$

Finally, an MNAR mechanism occurs if the missing observations depend on the specific value that is missing, that is, the missing value depends on the unobserved data [22].

Consider the following example, to illustrate each of the three missing mechanisms. Assume that a questionnaire to examine the relationship between obesity and sociodemographic variables was conducted and, among the questions, the household income of the participant was asked. If the percentage of missings for the income variable is higher among older participants than in younger participants, the missing values will likely depend on the observed values of another variable. In this situation, the missing mechanism is MAR because it suggests that the probability of the income answer being missing is related to the observed values of the variable age since older participants may be more hesitant to disclose their income. This results in a higher rate of missing data for that specific group. On the other hand, if data is missing due to data entry errors or due to the participant skipping a question without notice, the missing is MCAR because it is not related to either observed or unobserved values. These missing values occur randomly and do not show any systematic patterns or dependencies within the dataset. Usually, in a dataset with an MCAR mechanism, the percentage of missing values per variable is very similar among the variables, and these percentages are rather low. Lastly, if a participant refuses to answer because he/she has a low income, the missing depends on the unobserved value. In this case, participants with low income deliberately choose not to disclose their income information, so the missing mechanism is MNAR.

Another important aspect to consider when dealing with missing data is the percentage of missingness in the dataset. In fact, this has been one of the greatest issues among researchers since there is no defined value in the literature that establishes the proportion of missing data that allows for valid statistical inferences [25]. Despite this, the missing mechanism appears to be a more relevant question since several techniques to handle missing values depend on the missing pattern. Subsequently, I will present some methods commonly used when dealing with missing data.

2.1 Deletion Methods

Deletion methods are one of the most simple techniques to handle missing data. Listwise Deletion and Pairwise Deletion are the two main methods [35]. In listwise deletion, only complete cases are allowed, meaning that the entire vector of observations from an individual is removed if there is any variable with a missing value. Clearly, this procedure

may lead to a considerable decrease in the total sample size. Additionally, if the missing mechanism is not MCAR the analyses will produce biased results [35]. Similarly, pairwise deletion excludes an observation from the analysis of a specific variable if there is a missing value in that variable, but allows all the observations to be used for other variables with no missing values. As such, it allows for a larger sample size than listwise deletion, as it includes cases with missing values for some variables in analyses involving other variables. As before, pairwise deletion can lead to bias and inaccurate results if the missing mechanism is MNAR [35].

2.2 Maximum Likelihood Estimation

Maximum likelihood differs from the deletion methods presented since it makes use of the entire dataset, including the incomplete observations, to estimate the parameter values that maximize the probability of producing the sample data [35]. The parameters are estimated by maximizing the likelihood function. A maximum likelihood estimator (MLE) of a parameter θ has several desirable properties, including consistency (as the sample size n increases, the MLE converges in probability to the true value of the parameter θ), asymptotic normality (the distribution of the MLE becomes approximately normal as the sample size n increases), invariance (MLE is invariant to one-to-one transformations of the parameters) [26]. The method postulates a distribution and assumes at least, a MAR mechanism. If the missing mechanism is MNAR, this method may still yield biased estimates, although less biased than traditional techniques [35].

Let $Y_i = (Y_{i1}, ..., Y_{iK})$ be vector of observations for the experiment unit $i = 1, ..., n$ and θ the unknow parameters regarding the Y_i distribution. If there are no missing values, the log-likelihood is given by [24]:

$$l(\theta; Y_i) = \sum_{i=1}^{n} log f(Y_i; \theta) = \sum_{i=1}^{n} l_i(\theta; Y_i) \tag{2.3}$$

If the dataset contains missing values, the likelihood in equation (2.3) cannot be computed. The log-likelihood for the observed values is given by equation (2.4) considering O and R as defined in section 2 but with the indexes dropped for simplicity [24].

$$l_O(\theta; y_{iO}) = \sum_{i=1}^{n} log f(y_{iO}; \theta) = \sum_{i=1}^{n} l_{iO}(\theta; y_{iO}), \quad \text{where} \quad f(y_{iO}; \theta) = \int f(y_i; \theta)\, dy_{iR} \tag{2.4}$$

The marginal density of y_{iO} is computed by integrating out all possible values of the missing values (Y_{iR}). However, equation (2.4) can be too complex to compute so it might not be possible to maximize this log-likelihood function in order to obtain the estimates for the parameters. To overcome this obstacle, section 2.3 presents the Expectation-Maximization algorithm.

2.3 Expectation-Maximization

Expectation-Maximization (EM) is an iterative algorithm that can be used to maximize equation (2.4). It consists of two steps - Expectation (E-step) and Maximization (M-step) [36]. In the E-step, the algorithm replaces all missing values with their expected values, according to the current fit of the model. In the M-step, the algorithm re-estimates the parameters of the model using the filled-in data [24]. These two steps are repeated iteratively until the estimates of the model parameters converge [36]. Formally the method can be performed by following the next steps (where $\theta_{(t)}$ is the parameter value in the t-th iteration of the algorithm and $\theta_{(1)}$ is the starting value) [24]:

1. **E(Expectation)-step**

   Calculate
   $$Q(\theta;\theta_{(t)}) = \sum_{i=1}^{n} \int l_i(\theta) f(y_{iR}|y_{iO};\theta_{(t)})\, dy_{iR},$$
   where $y_i = (y_{iO}, y_{iR})$ and $l_i(\theta) = log f(y_i;\theta)$.

2. **M(Maximisation)-step**

   Maximize $Q(\theta;\theta_{(t)})$ with respect to θ, using its critical points:
   $$s(\theta_{(t+1)};\theta_{(t)}) = 0, \quad \text{where} \quad s(\theta,\theta_{(t)}) = \frac{\partial Q(\theta;\theta_{(t)})}{\partial \theta}$$

3. Return to Step 1 until convergence.

This method produces unbiased estimators when the data is MCAR and, for data with missing mechanism MAR, it provides less biased estimators than the traditional simple methods. However, the EM algorithm is not guaranteed to find the global maximum of $Q(\theta;\theta_{(t)})$ and it may be sensitive to the initial choice for the parameter estimates. In addition, the standard errors estimated by the method are lower than expected so some test statistics may be inaccurate [37].

2.4 Single Imputation

Single imputation is a simple technique to handle missing data in which a single value is used to replace each missing value in the dataset [35]. This replacement can be performed in several ways: (1) replacing the missing values by the mean (for continuous variables) or mode (for categorical variables) of that variable (Mean/Mode Imputation, it assumes the MAR mechanism), (2) replacing each missing values by a randomly selected value from the observed data (Random Imputation, it assumes the MAR mechanism), (3) the missing data in a given variable is estimated by a regression equation that uses other variables in the dataset (Regression Imputation, it assumes the MAR mechanism), (4) the missing value is estimated by a regression equation plus an additional residual (generated from a normal distribution with a mean of zero and a variance equal to the residual variance from the preceding regression analysis) (Stochastic Regression Imputation, it assumes the MAR or MCAR mechanism), (5) selecting a random observed value from a set of similar cases within the same classification group (Hot Deck, it assumes the MAR or MCAR mechanism) [35][26].

2.5 Multiple Imputation

Multiple Imputation (MI) was first proposed by Rubin in 1977, and is still one of the most used techniques to impute missing values [34]. In a simple way, this method generates multiple (say, m) complete versions of the dataset, each with potentially different imputed values for the missing data. These imputed datasets are then analyzed separately and the results are combined to provide a more accurate estimate of the parameter and its corresponding standard error. MI has an important advantage over single imputation since the generation of several imputed datasets accounts for the uncertainty associated with the imputed values. MI assumes data is MAR in order to achieve valid statistical inferences but it is also possible to apply it under an MNAR mechanism [22] [38]. MI consists essentially of two steps: Imputation and Pooling.

2.5.1 Imputation

The imputation step involves creating multiple versions of the dataset replacing missing values by estimated values based on the observed data. Imputation of multivariate data can be obtained essentially from two approaches: Joint Modeling (JM) ([39] [40]) or Fully

Conditional Specification (FCS) ([41] [42] [43]). Joint Modeling imputes missing values for each incomplete variable from a multivariate distribution, in a single step. The Fully Conditional Specification (FCS) involves imputing one variable at a time, based on the conditional densities of each incomplete variable, using the other variables in the model [28]. This approach is versatile, as different regression models can be used. For example, linear regression can be used for continuous variables and logistic regression can be used for categorical variables. FCS with MICE implementation is the most used method for multiple imputation and will therefore be presented in detail [13] [15] [19] [44].

The MICE algorithm assumes that the missing data mechanism is MAR and consists of the following steps [45][46][28]:

1. Single imputation (eg the Mean Imputation or Random Imputation) is performed for every missing value in the dataset.
2. For the vector representing one particular variable, say x_j, the values imputed in step 1 are set back to miss.
3. The observed values of the vector x_j are regressed on the remaining variables of the imputation model (it may consist of the entire dataset or of a subset).
4. The missing values for x_j are replaced by the model's predictions.
5. Steps 2-4 are repeated for every variable. This completes a cycle.
6. Steps 2-4 are repeated for a number of cycles established by the researcher.

At the end of each cycle, a complete dataset is obtained. The third step can be performed using several regression models and the model can differ according to the variable. For example, one can use MICE with random forests as the imputation model for each variable, or MICE with predictive mean matching, binary logistic regression models, and multinomial logistic regression models to impute continuous, binary, and categorical (with more than 2 levels) variables respectively. As Predictive Mean Matching is the default imputation model in R-package MICE, which we will use in our dataset, it will be explained throughout.

Predictive Mean Matching

As before, let x_j be a vector representing one particular variable, and Z be the set of the remaining variables in the dataset. Predictive mean matching is performed in the following way [42]:

1. Using the observed values in x_j, estimate a regression model (linear, logistic, or multinomial, depending on the variable) for x_j using Z.
2. Perform a random draw from the posterior predictive distribution of the set of coefficients (β) estimated in step 1. Let $\dot{\beta}$ be the new set of coefficients.
3. Predict all the entries of the vector x_j (even the ones without missing values).
4. For each observation of the x_j vector that initially has a missing value, identify a set of observations (called donors) of size d (previously defined by the user), with an observed value, whose predictive values from step 3. are similar to the predictive value for the observation with the missing value.
5. Randomly choose an observation from the set defined in 4. and replace the missing value by the observed value of that observation.

Let $x_{jobs} \in \mathbb{R}^{n_1}$ vector of observed data in the incomplete variable x_j, Z_{obs} be the $n_1 \times q$ matrix of predictors with rows corresponding to observed data in x_j, and Z_{mis} be the $n_0 \times q$ matrix of predictors with rows corresponding to missing data in x_j. The estimates for β and $\dot{\beta}$ are computed in the following way [42]:

- Calculate $S = Z_{obs}^t Z_{obs}$;
- Calculate $V = (S + diag(S)\kappa)^{-1}$, for small κ; the term $diag(S)\kappa$ is used to avoid problems with singular matrices;
- Estimate $\hat{\beta} = VZ_{obs}^t x_{jobs}$; this is an estimate of the coefficients in the regression linear model of x_{jobs} on the predictors of Z_{obs};
- Randomly draw a value g from a random variable $G \sim \chi^2_{n_1-q}$; $n_1 - q$ corresponds to the degrees of freedom in the model;
- Calculate $\dot{\sigma^2} = (x_{obs} - Z_{obs}\hat{\beta})^t (x_{obs} - Z_{obs}\hat{\beta})/g$; $\dot{\sigma^2}$ is the estimated residual variance in which the numerator corresponds to the residuals sum of squares (RSS);
- Randomly draw q independent $N(0,1)$ values and obtain a vector z_1;

- Compute $V^{1/2}$, using the Cholesky Decomposition;
- Estimate $\dot{\beta} = \hat{\beta} + \dot{\sigma} z_1 V^{1/2}$.

The method assumes that the distribution of the missing values is the same as that from the observed data of the donors. To execute the algorithm, it is essential to have a metric that quantifies the distance between observations, as well as a method for selecting a suitable donor. For example, let $\hat{x}_i$ denote the predicted value of the rows with an observed x_j and $\hat{x}_k$ the rows with a missing value for x_j. The donor selection could consider d candidates for which $|\hat{x}_i - \hat{x}_k|$ is minimal. The number of donors has to be specified by the researcher but typically it is set to 3, 5, or 10 [42]. The default value in the R-package MICE is 5, which will be used in this report.

Figure 2.1 shows a graphical representation of the predictive mean matching method; the colored orange rows represent the donors.

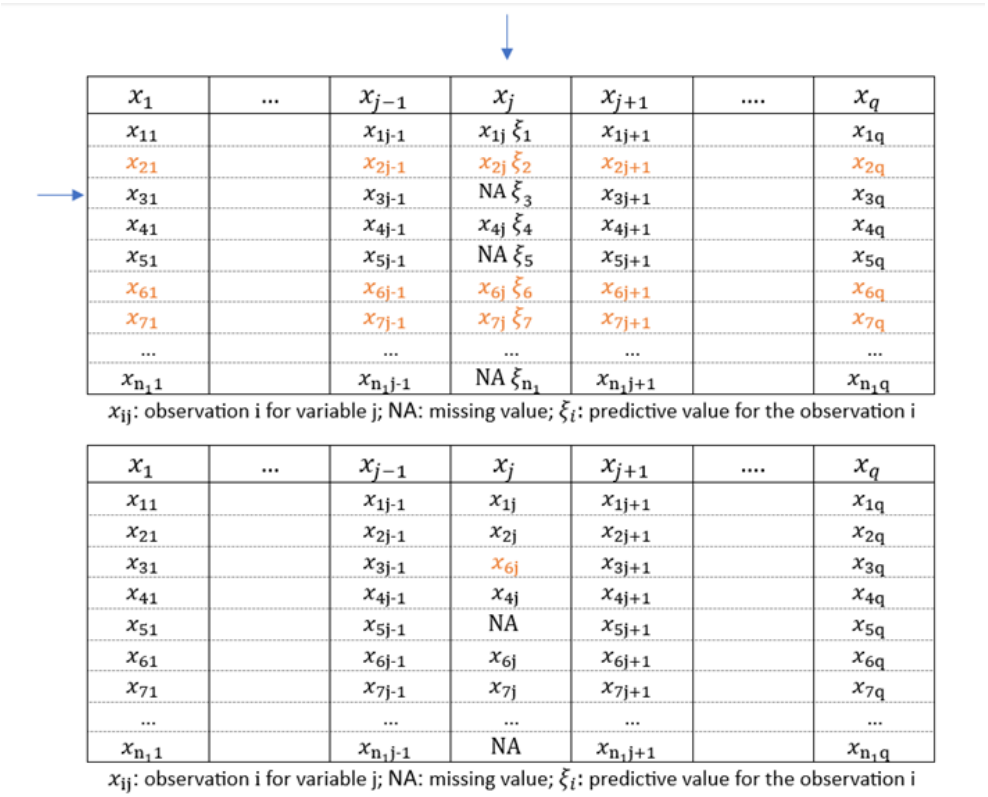


FIGURE 2.1: Example of the predictive mean matching method, representing the original dataset (tops) and the imputed dataset (bottom). The donors correspond to the colored rows.

2.5.2 Pooling

Pooling aims to combine the results from the separate analyses in order to provide a final estimate of the population parameters and standard errors. This can be done through several methods, but Rubin's rule is the most common in practice. Let $\hat{\theta}_i$ be an estimator of θ, obtained from the i^{th} dataset generated on cycle i (m cycles in total). Let $\hat{V}_i$ be its estimated variance. According to Rubin's Rule [39], the pooled point estimator of θ is calculated using equation:

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \tag{2.5}$$

When pooling the results, there are two variances to consider. The within-imputation variance ($\bar{V}$) and the between-imputation variance (B). The within-imputation variance refers to the variation of $\hat{\theta}$ within each imputed dataset and it is the average of the estimated variances in each of these imputed datasets [47]:

$$\bar{V} = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i$$

The between-imputation variance measures the variability of $\hat{\theta}$ between the imputed datasets and accounts for the extra variability due to the presence of missing values [47]:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2$$

The total variance of $\bar{\theta}$ is a weighted sum of the within-imputation variance ($\bar{V}$) and the between-imputation variance between variance (B) and is computed by:

$$Var(\bar{\theta}) = \bar{V} + (1 + \frac{1}{m})B \tag{2.6}$$

As equation (2.6) shows, the between-imputation variance has a slightly more significant weight in the value of the total variance. The addition of the factor $\frac{B}{m}$ allows accounting for the additional variability introduced by the fact that we are using a finite number of imputations instead of an infinite number [48][39]. Indeed, if $m = \infty$, $\frac{B}{m}$ would converge to 0 so the total variance would be just the sum of the between- and within-imputation variances. Additionally, increasing the number of imputations can lead to greater efficiency and more accurate estimates if the within-imputation variance is small compared

to the between-imputation variance. On the other hand, if the within-imputation variance is large, compared to the between-imputation variance, increasing the number of imputations may not result in substantial improvements in accuracy [39].

Another aspect to consider when pooling the results is the number of degrees of freedom of the pooled t-test. The computation of the degrees of freedom was initially proposed by Rubin, in 1987 [39]. The value was computed as shown below, where λ is the proportion of the total variance that is due to missingness:

$$df_{old} = \frac{m-1}{\lambda^2}, \quad \text{where} \quad \lambda = \frac{B + \frac{B}{m}}{Var(\bar{\theta})}$$

This number of pooled degrees of freedom is larger than the number of degrees of freedom of each imputed dataset, which is not accurate. Therefore, Barnard and Rubin defined the number of adjusted pooled by [49]:

$$df = \frac{df_{old} * df_{observed}}{df_{old} + df_{observed}}, \quad \text{where} \quad df_{observed} = \frac{(n-k)+1}{(n-k)+3} \times (n-k)(1-\lambda)$$

With all the previous knowledge, it is now trivial to construct a $1-\alpha$ confidence interval for θ, namely:

$$\bar{\theta} \pm t_{df,1-\frac{\alpha}{2}} \sqrt{Var(\bar{\theta})}$$

2.5.3 Efficiency

There are essentially two questions that should be addressed when performing multiple imputation. What is the maximum percentage of missing values that is allowed in order to obtain good inference results? How many imputed datasets should be considered?

The first question is seriously relevant since most techniques evidence a decrease in efficiency and quality when the missing rate is very high. Despite this, there is no agreement in the statistics community about the adequate percentage. Several papers have conducted simulation studies using either real or simulated datasets, exploring various percentages of missing data ranging from 10% to 80% under the three missing mechanisms (MCAR, MAR, MNAR) [25] [26] [27] [50]. The key findings across the studies are that the bias of the estimated parameters increases when the rate of missing information increases. Despite this, the bias and standard deviations varied a lot, depending on the missing mechanisms. A recent study stated that the missing percentage should not guide

decisions on multiple imputation since the results vary, depending on: (1) the missing mechanism; (2) the role played by the variables with missing values (outcome, exposure, or confounder); (3) the sample size; (4) the inclusion of auxiliary variables on the imputation model [51]. This led to the conclusion that it is not adequate to establish a cutoff value for the percentage of missing and that every scenario should be analyzed carefully.

The second question relies on the concept of efficiency. It is intuitive that a dataset with a large percentage of missing values requires a larger number of imputations than otherwise [38]. Rubin provides an equation for the Relative Efficiency (RE) based on the number of imputed datasets (m) and the fraction of missing information (γ). RE gives information about the accuracy of the parameter estimate [39]. If RE is close to 1, it suggests that the imputation method has preserved the same level of accuracy in estimating the parameter as if there were no missing data.

The equation for RE is given by:

$$RE = (1 + \frac{\gamma}{m})^{-1}, \quad \text{where} \quad \gamma = \frac{RIV + \frac{2}{df+3}}{1 + RIV} \tag{2.7}$$

there, RIV is the Relative Increase in Variance due to nonresponse: it is also an important measure when dealing with missing data. It corresponds to:

$$RIV = \frac{B + \frac{B}{m}}{\bar{V}} \tag{2.8}$$

For example, if a variable contains 50% of missing information (γ=0.5) and 10 imputed datasets were considered, the method achieves a relative efficiency of 95%. Despite this formalism, there is no consensus about the adequate value of m since it depends on the value of the efficiency that we are trying to achieve [25].

A question that can now arise is how similar are the m imputed datasets? To approach this question, agreement measures will be computed (as in Chapter 3). Table 2.1, adapted from [52], shows those agreement measures according to the type of variable under analysis. Fleiss'kappa and the intra-class correlation coefficient will be used since m is usually greater than 2.

Fleiss'Kappa

This measure quantifies the achieved agreement beyond chance as a proportion of the potential agreement beyond chance [53]. The index is computed using equation (2.9),

TABLE 2.1: Methods for assessment of the agreement between values (adapted from [52])

Type of variable	Number of observers between whom agreement is to be assessed	Method for assessing agreement
Categorical (nominal)	2	Cohen's kappa
	> 2	Fleiss' kappa
Categorical (ordinal)	2	Weighted kappa
	> 2	Fleiss' kappa
Continuous	Two or more observers or techniques	Intra-class coefficient Bland-Altman plot with limits of agreement

where $\bar{P}$ is the proportion of observed agreements and $\bar{P}_e$ is the proportion of agreement expected by chance [53] [54].

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2.9}$$

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right)$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \qquad\qquad p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$$

In the previous equations, N represents the total number of subjects, n is the number of ratings per subject, and k represents the number of categories available for assignments. The subscript i takes values from 1 up to N, representing the individual subjects being rated. The subscript j takes values from 1 up to k, representing the categories into which the assignments are made. Finally, n_{ij} is the number of raters who assigned the i^{th} subject to the j^{th} category.

Kappa values range from 0 to 1, where 0 represents no agreement beyond chance, and 1 represents perfect agreement. Negative kappa values are also possible but uncommon since they only occur when the observed agreement among raters is worse than what would be expected by chance alone.

Intra-class correlation

The Intra-Class Correlation (ICC) coefficient is calculated as the ratio of the between-group variability to the total variability (equation (2.10)), thus corresponding to the proportion of the total variability that is due to the between-group variability. The ICC is also equal to the correlation between the repeated measures of each observation, (if such

a setting exists), linear mixed model as described in equation (2.11). The coefficient ranges from 0 to 1, with 1 meaning excellent agreement.

$$ICC = \frac{\sigma_0^2}{\sigma_0^2+\sigma^2} = Corr(y_{ij}, y_{is}) \quad (j \neq s) \tag{2.10}$$

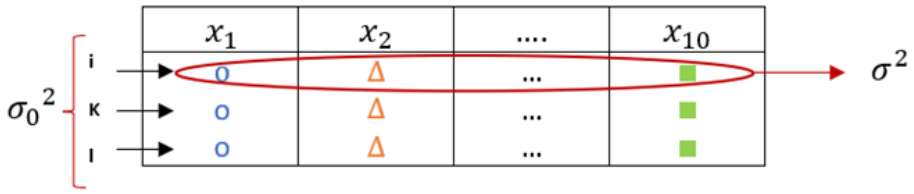


FIGURE 2.2: Example of a dataset that illustrates the variability within observations and between observations.

Subject i, Replicate j:

$$Y_{ij} = \beta_0 + b_{0i} + u_{ij}\,, \quad \text{with} \quad b_{0i} \sim N(0, \sigma_0^2)\,, \quad u_{ij} \sim N(0, \sigma^2) \tag{2.11}$$

2.6 R-Packages

All the statistical analyses carried out within the scope of this report were carried out in R. In what follows, I present the R-packages most commonly used to handle missing values.

MICE

This package is one of the most used to impute missing values. The package allows for multiple imputation using Fully Conditional Specification (FCS) by implementing the MICE algorithm described in section 2.5.1 [28] [55]. The package allows users to specify different imputation models for different types of variables, such as continuous, binary, categorical, or ordinal. It provides various functions to explore, compare and evaluate the quality of the imputed data, as well as to pool the results from multiple imputations for statistical inference by Rubin's rule. MICE assumes the data is missing at random (MAR).

Amelia II

This package also performs multiple imputation and implements the EMB (Expectation-Maximization with Bootstrapping) which makes it faster and more robust to impute many variables including for both cross-sectional and time series data [56]. The algorithm utilizes the EM (expectation-maximization) algorithm on multiple bootstrapped samples of

the original incomplete data to estimate the complete-data parameters. Subsequently, imputed values are drawn from each set of bootstrapped parameters and used to replace the missing values in the dataset. Amelia II has two important assumptions: 1) the entire dataset should be multivariate normal, and 2) the missing mechanism is MAR [56]. The first assumption is often unverified but transformations to achieve normality can be performed in advance.

missForest

The missForest package implements a nonparametric imputation method. It uses a random forest trained on the observed values of the dataset to predict the missing values. The method works for continuous and/or categorical data including interactions and non-linear relations [57]. The missForest package offers a method for estimating imputation error without relying on a separate test set or extensive cross-validation.

Multiple imputation with diagnostics (mi)

The mi package provides functions for performing multiple imputation. It offers various methods, including predictive mean matching, regression imputation, and random forests imputation, among others [58]. It also provides tools for analyzing and summarizing the results of multiple imputation, such as combining the imputed datasets by Rubin's rules and performing imputation diagnostics.

Chapter 3

Variable Selection

The primary objective of the internship was to develop a predictive model for childhood obesity at the age of 13. Thus, this chapter introduces various variable selection methods that will be employed to select the final model. Variable selection methods are essential to identify the subset of variables that best explain a given response variable. In the context of having too many explanatory variables, the inclusion of irrelevant or redundant variables can lead to complex models, overfitting problems, and poor model performance. Thus, selecting only the predictors that have a relevant association with the response variable is imperative. Feature selection methods on a labeled dataset (supervised methods) can be divided into three categories: filter models, wrapper models, and embedded models, and are presented in the following sections [29].

3.1 Filter Models

In filter methods, the selection of variables occurs without any machine learning classification algorithm [29]. These methods select the features based on two steps: 1) the features are scored based on certain criteria, and 2) the features with the highest scores are selected [29]. Some possible criteria to use in Step 1 are, for example, correlation metrics (Pearson, Spearman), Chi-squared tests, Fisher's Score, and ANOVA. Additionally, step 1 can be conducted either in a univariate or multivariate setting [29]. Under the univariate scheme, each feature is evaluated independently of the other features in the dataset. Univariate filter methods are computationally efficient and easy to interpret, but they may not capture interactions or dependencies among the features. In contrast, the multivariate scheme evaluates features in a batch, meaning that it evaluates the entire feature space

as a single entity, and not as a collection of individual features. This allows for the capture of interactions and dependencies among the features and for the evaluation of their collective impact on the response variable. Although simple, these methods have a great disadvantage, as they ignore the effects that the selected variables have on the model performance [29]. So, even if the method chooses the most appropriate variables, based on specified criteria, these variables may result in a model with poor predictive performance. To overcome this limitation, the following two types of models were developed.

3.2 Wrapper Models

Wrapper models select a subset of variables to train the model, and based on the inferences and performance of that model, decide to add or remove a feature from the initial subset [29]. These methods involve three steps: 1) searching a subset of variables through the space of all possible subsets of features, 2) evaluating the selected subset of variables by the performance of the classifier, and 3) repeating steps 1 and 2 until the desired performance is reached or until a satisfactory subset of features is selected [29]. To perform step 1, it is necessary to choose a technique for searching for the subset of variables [29]:

- Hill-climbing starts with a small set of features and gradually expands the subset to the subset with the best performance. This process is repeated until no further improvements in performance are observed.
- Foward selection begins with an empty set of variables and, in each iteration, the most relevant predictor is added to the set. The relevance of each variable is defined by the user but is based on an improvement of the model's performance.
- Backward elimination starts with the full set of variables and then iteratively removes the least relevant feature until a desired level of model performance is achieved. At each iteration, a variable is removed and the model is re-fitted to choose the new variable to be excluded. Similar to forward selection, the relevance of each variable is based on an improvement of the model's performance.
- Stepwise Selection is a combination of backward elimination and forward selection since in each step, the method iteratively adds or removes one variable at a time, based on pre-specified criteria (Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), or the p-value of the predictor variable).

- Boruta algorithm [59]: The algorithm starts by duplicating the dataset under analysis and shuffling the values in each column. The new columns are called shadow features and are used as a reference to determine whether a variable is important or not. On this new dataset containing the initial and shadow features, a classifier such as a Random Forest Classifier is used to train the model concerning the outcome variable. Then, a measure such as the Mean Decrease Accuracy or the Mean Decrease Impurity is computed to evaluate the importance of each feature of the dataset. The original features that have a higher importance score than the best of the shadow features (i.e. the features that have a higher Z score than the maximum Z score among the shadow features) are marked as a "hit" and are relevant for the analysis. The algorithm repeats this process several times, each time with a new set of randomly permuted shadow features. This helps to ensure that the results are not biased by the initial set of shadow features. A feature is rejected and excluded from the collection of predictors if the feature has not been marked as a hit in a specified number of iterations. The algorithm stops after several of iterations (for example if it reaches the maximum number of iterations of the random forest) or if all features have been confirmed or rejected. Figure 3.1, adapted from [60], shows a visual demonstration of the Boruta algorithm.

Wrapper models result in better predictive accuracy estimates than filter models but are more demanding computationally.

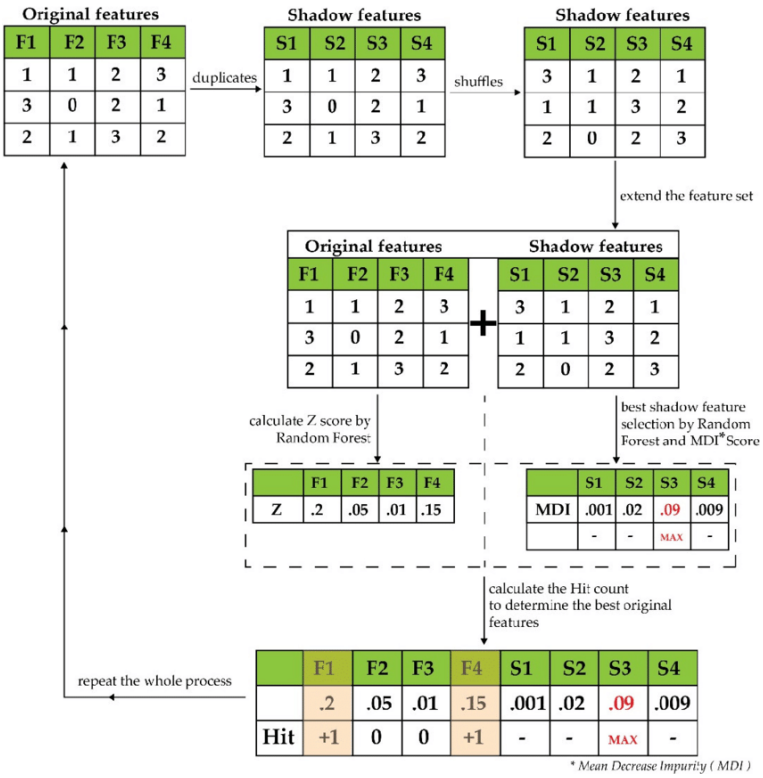


FIGURE 3.1: Boruta algorithm (adapted from [60]).

3.3 Embedded Models

Embedded models combine feature selection and model fitting in a single step. These algorithms select the most relevant variables while training the model, rather than performing feature selection as a separate step before model training [29]. Embedded Models are the best of both worlds regarding wrapper models and filter models since, similar to the former, they include the interaction with the classification model when choosing the variables, but are far less computationally demanding [29]. Regulation models such as Least Absolute Shrinkage and Selection Operator (LASSO) [61], Adaptative LASSO (ALASSO) [62], Ridge [63], Elastic Net (ENET) [64], and Adaptative Elastic Net (AENET) [65] are some of the examples of most used embedded models. In these methods, the parameters are obtained by minimizing the log-likelihood function of the data with the addition of a penalty function that differs according to the method.

$$\hat{\theta} = \arg min_{\theta}\{-\frac{1}{n}\sum_{i=1}^{n} logL(\theta|Y_i, X_i) + \lambda P_{\alpha}(\beta)\} \tag{3.1}$$

Here, L is the likelihood function, β is the $p \times 1$ vector of regression coefficients and $\theta = (\beta_0, \beta)$ is the total vector of the regression parameters which includes the intercept. The

penalty function $P_\alpha(\beta)$ varies according to the penalized regression under analysis:

1. Ridge: $P_\alpha(\beta) = \sum_{j=1}^{p} \beta_j^2$
2. LASSO: $P_\alpha(\beta) = \sum_{j=1}^{p} |\beta_j|$
3. ALASSO: $P_\alpha(\beta) = \sum_{j=1}^{p} \hat{a}_j |\beta_j|$
4. ENET: $P_\alpha(\beta) = \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2$
5. AENET: $P_\alpha(\beta) = \alpha \sum_{j=1}^{p} \hat{a}_j |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2$

Above, β_j is the regression coefficient for the j-th variable in the dataset. The adaptive weights ($\hat{a}_j$) allow for different penalized values for each variable.

In Ridge regression, the coefficients are shrunk towards 0 but the method does not force them to be exactly zero. Ridge regression is useful when there are many correlated predictors, as it can reduce the impact of multicollinearity on the coefficients. In Lasso regression, the penalty function is the sum of the absolute values of the coefficients resulting in a sparse model. Therefore, many of the coefficients will be exactly zero. Lasso is useful when there are many predictors, but only a few of them are expected to be important. Elastic Net is a combination of Ridge and Lasso regression, where the penalty term is a mixture of the sum of the squared coefficients and the sum of the absolute values of the coefficients. Elastic Net can handle many correlated predictors; it still performs variable selection by shrinking some of the coefficients to zero. Lastly, the adaptive versions of the methods allow the incorporation of adaptive weights to the penalty term, which depend on the magnitude of the coefficients estimated in the previous iteration. The adaptive weights essentially down-weight the importance of the variables that are less important and up-weight the importance of the variables that are most important.

3.4 Variable selection on multiply imputed datasets

The methods presented in the previous sections are easy to implement in the presence of a dataset without missing values. However, a problem arises when the dataset contains missing values and the multiple imputation method is used to complete the dataset. This is a serious question to address because, from the MICE procedure, m different datasets are generated and a different set of variables may be selected in each imputed dataset. To

avoid this, I present five procedures that allow us to perform variable selection considering multiply imputed datasets [66]:

1. Fit the model on all imputed datasets and use Rubin's Rule to combine the estimates for the coefficients and corresponding p-values of the Wald test. Then, perform backward selection or forward selection.

2. Select the interesting variables in the subset of the dataset that contains only complete observations (observations without missing values) and then fit the model with the selected variables on all imputed datasets. This technique may produce biased results under the MAR missing mechanism.

3. Use single stochastic imputation and randomly choose one of the m imputed datasets to fit the model and perform variable selection. This method produces biased estimates and a different model may be selected depending on the imputed dataset chosen.

4. Restrict the exposures that were selected: 1) in any model (this can result in several selected variables), 2) in at least half of the models, 3) in all models (can result in a small set of variables).

5. Stack the m imputed datasets and create one large dataset of length $m \times n$. By fitting the model to this new dataset, the parameter estimates are valid but the standard errors are too small, which is inaccurate [66]. However, we can correct the standard errors and apply a fixed weight (w_i) to all observations:

   - $w_i = \frac{1}{m}$. By using this weight the log-likelihood for the stacked dataset is scaled to the equivalent of a dataset of length n (as the initial dataset).
   - $w_i = \frac{(1-f)}{m}$, where f is the average fraction of missing data across all variables. f is calculated by considering the total number of missing values across all variables divided by pn.
   - $w_i = \frac{(1-f_i)}{m}$, where f_i is the fraction of missing data for variable x_i. f_i is calculated by computing the number of missing values for variable X_i divided by n.

The previous methods can be considered relatively straightforward. Alternatively, a more advanced approach for variable selection involves utilizing penalized regression

methods such as LASSO, ALASSO, Ridge, ENET, and AENET. However, applying a variable selection algorithm on every imputed dataset is likely to result in distinct sets of chosen predictors. If we fit a penalized regression method on each imputed dataset we want to minimize the following function [67]:

$$\hat{\theta}_d = \arg min_{\theta_d}\{-\frac{1}{n}\sum_{i=1}^{n} logL(\theta_d|Y_{d,i}, X_{d,i}) + \lambda P_\alpha(\beta_d)\} \tag{3.2}$$

Here, L is the likelihood function, β_d is the $p \times 1$ vector of the regression coefficients for the d-th dataset and $\theta_d = (\beta_{0_d}, \beta_d)$ is the total vector of the regression parameters which includes the intercept, and the subscript d takes values from 1 up to m, $P_\alpha(\beta_d)$ is the penalty function that was defined in 3.3.

A recent study considers a different penalized function by pooling the objective functions across imputations, which forces the algorithm to select the same set of variables across the imputed datasets [67]. The paper proposes two different ways to construct the final objective function: the stacked function and the grouped function.

Stacked function

To form the stacked objective function, we consider the sum of the loss functions for each imputed dataset and jointly optimize the collective objective function:

$$\hat{\theta} = \arg min_{\theta}\{-\frac{1}{n}\sum_{d=1}^{m}\sum_{i=1}^{n} logL(\theta|Y_{d,i}, X_{d,i}) + \lambda P_\alpha(\beta)\} \tag{3.3}$$

Looking at equation (3.3), and in contrast with equation (3.2), we see that θ is not indexed by d which implies that the optimization of equation (3.3) will result in one estimated parameter vector $\hat{\theta}$. Thus, the same variables will be selected across all imputed datasets. However as described in the fifth rule above (5), stacking all imputed datasets may be seen as a way to inflate the effective sample size. So we can consider a weight for each subject (w_i) and the collective objective function becomes [67]:

$$\hat{\theta} = \arg min_{\theta}\{-\frac{1}{n}\sum_{d=1}^{m}\sum_{i=1}^{n} w_i logL(\theta|Y_{d,i}, X_{d,i}) + \lambda P_\alpha(\beta)\} \tag{3.4}$$

The penalty function $P_\alpha(\beta)$ does not depend on the imputed dataset d, and it is defined as in 3.3.

Grouped function

The grouped objective function adds a group LASSO penalty across the imputed datasets [67]:

$$(\hat{\theta}_1, ..., \hat{\theta}_m) = \arg min_{\theta_1,...,\theta_m}\{-\frac{1}{n}\sum_{d=1}^{m}\sum_{i=1}^{n} logL(\theta_d|Y_{d,i}, X_{d,i}) + \lambda P(\beta_1, \beta_2, ..., \beta_m)\} \quad (3.5)$$

Chen and Wang [68] formulated a special case for the grouped function with the penalty function given below, forming the MI-LASSO method:

$$P(\beta_1, \beta_2, ..., \beta_m) = \sum_{j=1}^{p}\sqrt{\sum_{d=1}^{m}\beta_{d,j}^2}.$$

Equation (3.5) shows that the parameter vector (θ) is indexed by d which means it may be different across the imputed datasets. Despite this, for any fixed j the group LASSO penalty jointly shrinks all $\beta_{d,j}$'s to zero so $\beta_{1,j} = ... = \beta_{m,j} = 0$ which allows the selection of the same variables across the imputed datasets. The following penalty functions are considered [68]:

- LASSO: $P(\beta_1, \beta_2, ..., \beta_m) = \sum_{j=1}^{p}\sqrt{\sum_{d=1}^{m}\beta_{d,j}^2}$
- ALASSO: $P(\beta_1, \beta_2, ..., \beta_m) = \sum_{j=1}^{p}\hat{a}_j\sqrt{\sum_{d=1}^{m}\beta_{d,j}^2}$

More details about the optimization process and tuning parameters for both functions can be found in [67].

Chapter 4

Methods for Modeling a Static Outcome with Longitudinal Predictors

This study aims to develop a predictive model for childhood obesity at age 13 considering exposures collected at several time points. In particular, the model will include longitudinal predictors (predictors measured throughout time). However, traditional longitudinal models are applied for a longitudinal response variable, and in our setting, that response variable is static. There are simple alternatives that allow us to incorporate longitudinal predictors while modeling a static outcome [31] [32] [33]:

1. To consider each measure as a predictor. However, treating each exposure as a separate predictor can lead to a significant increase in the number of predictor variables which can increase the complexity of the problem. Moreover, since the predictors were collected at different time points, they are most likely correlated which may easily lead to multicollinearity issues.

2. To select, for each longitudinal predictor, its values at the time point that best represents the predictor and/or is mostly correlated with the response. For example, the most important time point to predict childhood obesity will probably be the one that is closest to the time point of the outcome.

3. To consider summary statistics, like the mean or the maximum value, for the measurements of a single predictor in different time points, as predictors. Similarly to the strategy presented in item 2, we would lose the longitudinal information.

4. To consider each measurement at a specific time point, regress all previous measurements on it, and use the residuals on the regression model (conditional measures). The model should include the first or last measurements and all the subsequent conditional measures. Although it can be relevant to identify the periods mostly related to childhood obesity, this method will also lead to multicollinearity issues.

5. To perform a linear (mixed-effects) regression analysis with the time point at which the measurements were made as the independent variable and the longitudinal predictor as the dependent variable. Then the intercept and the slope of the regression enter the final model. Alternatively, we can fit a mixed-effects regression and the final model will include the random intercept and random slope. This regression can be formulated as defined in equation (4.1). There, a_{0i} and a_{1i} are the random intercept and random slope for individual i, respectively, and are jointly distributed as bivariate normal random variables, β_0, β_1, and β_2 are fixed effects, e_{ij} is the error term for individual i measured at time j, X_{ij} is a longitudinal exposure for subject i measured at time j, Y_i is the (cross-sectional) response variable for individual i, and t_{ij} is the time of measurement.

$$X_{ij} = a_{0i} + \beta_0 + a_{1i}t_{ij} + \beta_1 t_{ij} + e_{ij} \tag{4.1}$$

$$logit(P(Y_i|a_{0i}, a_{1i})) = \beta_0 + \beta_1 \hat{a}_{0i} + \beta_2 \hat{a}_{1i} \tag{4.2}$$

This approach is relevant if it is reasonable to assume that the predicted individual-specific intercepts and slopes for X accurately capture distinct patterns across the two outcome groups.

Another possibility that will also be explored in this report is to use regularization models like penalized logistic regression with LASSO. This is an alternative, as regularization methods are effective in handling multicollinearity issues, and all exposures can be included in the model.

Finally, I present finite mixtures of regression models as a different alternative. The idea is to extend LASSO regression, in order to include a finite number of different values

for the regression coefficients, thus allowing for the existence of a finite number of regressions in the data. In essence, we will allow for the existence of a finite number of (disjoint) clusters, and, on each cluster, a (most likely, different) LASSO regression model will be fitted, all at once.

A finite mixture model is a statistical model that assumes that a population is composed of multiple latent subpopulations or components, each following a different probability distribution [69] [70]. These subpopulations are typically assumed to be drawn from a finite set of distributions, such as Gaussian, Poisson, or exponential distributions. A finite mixture of regressions is a finite mixture such that on each component, the modeling of the mean has its own regression model. Therefore, by fitting a finite mixture of regression, it is possible to have distinct regression models for each subpopulation [70]. If there are indeed several components in the data, this allows each individual to be predicted by a most suitable regression. For K components, a finite mixture of regressions is defined by:

$$h(y|x,\phi) = \sum_{k=1}^{K} \pi_k f(y|x,\theta_k) \qquad \pi_k > 0 \qquad \sum_{k=1}^{K} \pi_k = 1 \tag{4.3}$$

where y is the dependent variable with conditional density h, x is the vector of predictors, π_k is the prior probability of component k, θ_k is the vector of the component-specific parameters (modeling mean or variance, for example) for the density function f, and $\phi = (\pi_1, ...\pi_k, \theta_1^T, ..., \theta_K^T)^T$ is the vector containing all parameters. In its simplest form, y is a univariate normal density, $\theta_k = (\mu_k, \sigma_k^2)$ and equation (4.3) defines a mixture of standard linear regression models (latent class regression) [71].

It is also relevant to define the posterior probability for an observation (x, y) belonging to class j, equation (4.4). Once the posterior probabilities are computed, each observation is assigned to the component with the highest posterior probability. This means that each observation has a certain probability of belonging to each component, but ultimately each observation is assigned to a unique component with its own regression model.

$$P(j|x,y,\phi) = \frac{\pi_j f(y|x,\theta_j)}{\sum_k \pi_k f(y|x,\theta_k)} \tag{4.4}$$

The unknown parameters (ϕ) of the mixture model include the mixture proportions and the parameters of the component distributions. These parameters can be estimated using statistical techniques such as maximum likelihood estimation or Bayesian inference [69].

We shall focus on the maximum likelihood estimation since we will be using an R-package that implements it.

The log-likelihood of a finite mixture of regressions given a sample of N observations is given by [70]:

$$logL = \sum_{n=1}^{N} log(h(y_n|x_n, \phi)) = \sum_{n=1}^{N} log(\sum_{k=1}^{K} \pi_k f(y_n|x_n, \theta_k)) \tag{4.5}$$

The maximum likelihood estimation includes maximizing equation (4.5) to estimate $(\pi_k, \theta_k)_k$, by making use of the EM algorithm [72]. The algorithm is divided into two steps [70]:

Step E: For each observation, estimate the posterior class probabilities using equation (4.4) and then compute the prior class probabilities $\hat{\pi}_k$.

$$\hat{p}_{nk} = P(k|x_n, y_n, \hat{\phi}) \qquad\qquad \hat{\pi} = \frac{1}{N} \sum_{n=1}^{N} \hat{p}_{nk}$$

Step M: Maximize the log-likelihood for each component separately using the posterior probabilities as weights:

$$max_{\theta_k} \sum_{n=1}^{N} \hat{p}_{nk} log(f(y_n, |x_n, \theta_k))$$

The E- and M-steps are iteratively performed until either the improvement in likelihood falls below a predetermined threshold or a maximum number of iterations is reached. This iterative process ensures that the model's parameters are continually updated to optimize the likelihood of the observed data.

The implementation of the model with the EM algorithm can be done in R with the package *flexmix* [70] [73]. This package allows fitting a finite mixture of regressions with various probability distributions, including Gaussian, Poisson, and Binomial, and also allows fitting mixtures of general linear models with penalization, like a LASSO or elastic net. It supports both continuous and categorical response variables.

Chapter 5

Dataset

The analysis will be conducted on the study population from the Generation XXI (G21) cohort. This is a population-based birth cohort in the Porto Metropolitan Area that, between April 2005 and August 2006, recruited 8647 newborns and their mothers from the five level-III public maternities in this area [21]. The recruitment procedure included mothers living in one of the six municipalities of the metropolitan area of Porto that gave birth to live babies with gestational age $>$ 24 weeks at the public hospitals in that area. Among the women invited to participate, 91% agreed to engage. The study participants were then followed at 4,7,10 and 13 years old (y.o.), with collected data on demographic and psychosocial factors, lifestyle behaviors, and anthropometric measures. The participation rates for each follow-up were 86 %, 80 %, 74 %, and 53 % respectively. The University of Porto Medical School/S. João Hospital Centre Ethics Committee approved the study and all the procedures regarding data collecting. In addition, all the participants signed an informed consent according to the Helsinki Declaration [74]. Although the initial dataset contained 8647 newborns, some were excluded from the analysis. The first eligibility criterion was to include only observations with a recorded value on the outcome variable (obesity at age 13), resulting in a sample size reduction by almost half compared to the initial count. This decrease was primarily attributed to the evaluation at age 13 taking place during the COVID-19 pandemic. We also excluded twins, newborns with congenital malformations, and underweight children. The final dataset included 4246 mothers-child pair.

5.1 Exposures

This report incorporates a set of 55 variables that were assessed during pregnancy, infancy, and childhood at ages 4,7,10. These exposures include sociodemographic and psychosocial factors, anthropometric measures, and lifestyle behaviors. The questionnaires utilized to gather data for the dataset slightly vary from one follow-up to another, resulting in certain variables not being available at each time point. Figure 5.1 shows the available exposures for the pregnancy and childhood stages according to each exposure group.

Exposure Group	Pregnancy/Infancy	Childhood 4 years	Childhood 7 years	Childhood 10 years
Sociodemographic and Psychosocial exposures	Maternal age, Parity, Maternal education level, Maternal working condition, Marital Status, Household income, Child sex.	Maternal education level, Maternal working condition, Marital Status, Household income.	Maternal education level, Maternal working condition, Marital Status, Household income.	Maternal education level, Household income.
Anthropometric exposures	Pre-conception Body Mass Index (BMI), Weight gain during pregnancy, Newborn's weight, Newborn's length, BMI at 6 months, 1 year and 2 years.	Child BMI z-score.	Maternal BMI, Child BMI z-score.	Child BMI z-score.
Lifestyle behaviors	Maternal Smoking habits during pregnancy (Y/N), First solid food introduced, Age of first solid food introduced, Breastfeeding.	Child sedentary time (hours), Child physical activity (hours), Child active play (hours), Child sleep duration (hours), Soup, Vegetables, and fruit consumption (portions per day), Soft drinks consumption, Calorie's consumption (Kcal).	Child sedentary time (hours), Child physical activity (hours), Child active play (hours), Soup, Vegetables, and fruit consumption (portions per day), Soft drinks consumption, Calorie's consumption (Kcal).	Child sedentary time (hours), Child physical activity (hours), Soup, Vegetables, and fruit consumption (portions per day), Soft drinks consumption, Calorie's consumption (Kcal).
Clinical exposures	Gestational diabetes (Y/N), Gestational hypertensive complications (Y/N), Mode of delivery (vaginal, Y/N), Mode of delivery (caesarean, Y/N), Gestational weeks.			

FIGURE 5.1: Exposures of the study.

Sociodemographic and Psychosocial exposures

Sociodemographic and psychosocial factors include maternal age at delivery (years); parity (nullipara or multipara); the educational level of the mother that was initially measured in years but was later categorized into three categories: basic ($\leq$ 9 years of education), secondary (10–12 years), and tertiary ($\geq$ 13) which corresponds to the education in Portugal based on the International Standard Classification of Education – ISCED; maternal marital status (married/cohabiting or no partner); maternal working condition (paid job or no paid job activity); household income Euros/month (low $\leq$1000; middle 1001–1500; high $>$ 1500); child's sex (female; male). All exposures are available at pregnancy, 4 y.o., and 7 y.o., except for parity, which was collected at birth. The 10-year follow-up includes only the exposures related to the mother's education and household income.

Anthropometric exposures

Anthropometric exposures regarding the pregnancy stage and the first 2 years of the child include the pre-conception Body Mass Index (BMI) which was categorized according to the WHO recommendations for adults ($<$25-underweight/normal or $\geq$ 25- overweight/obese); weight gain during pregnancy that was measured in kilograms but was also categorized according to 2009 IOM guidelines 12.5-18 kg, 11.5–16 kg, 7–11.5 kg, and 5–9 kg for women with pre-pregnancy BMI classified as underweight ($<$18.5 kg/m^2), normal weight (18.5–24.9 kg/m^2); overweight (25–29.9 kg/m^2) and obese ($\geq$ 30 kg/m^2) respectively; newborn's weight (g) and length (cm); BMI of the child collected at 6 months, 1 year and 2 years old. The mother's BMI was collected again at the 7-year follow-up. The child's BMI was calculated at 4, 7, and 10 years, and age and sex-specific BMI standard deviation scores (BMI z-scores) were derived based on the World Health Organization (WHO) growth reference values [75]. The BMI z-score was categorized as follows: normal weight ($>$ -2 SD and $<$1 SD) and overweight/obesity ($\geq$1 SD).

Lifestyle behaviors

Lifestyle behaviors regarding the pregnancy stage and the first 2 years of life include maternal smoking habits during pregnancy (yes or no); first solid food introduced (cereal porridge or fruit or soup); the age of first solid food introduced (before 4 months or between 4 and 5 months or after 6 months); breastfeeding (never or until 0-2.9 months or until 3-5.9 months or more than 6 months). Regarding the preschool and school years (4,7,10 y.o.), the exposures included are child sedentary time (hours); child physical activity (hours); child active play (hours); child sleep duration (hours); soft drinks consumption (never/less than 1 per week or 1-6 per week or one/more per day); the consumption

of soup, vegetables, and fruit that were assed simultaneously by summing up the values of the 3 exposures and considering the WHO recommendation of 5 portions a day (less than 5 a day or $\geq$ to 5 a day); calorie's consumption (Kcal). The child's sleep duration was only available at age 4 and the hours of active play were not available at age 10. All the remaining variables were collected at each time point (4,7,10).

Clinical exposures

Clinical exposures refer only to the pregnancy stage and include gestational diabetes (yes or no); hypertensive complications that were created based on the information of gestational hypertension, pre-eclampsia, or eclampsia in the following way: if the mother reported at least one of the mentioned complications, the variable's value was yes, and no otherwise; vaginal as the mode of delivery (yes or no); cesarean as the mode of delivery (yes or no); the number of gestational weeks.

Outcome

For each participant, the body height was measured standing upright using a stadiometer (Seca 218, Seca Corporation, Hamburg, Germany), to the nearest 0.1 cm. The participant's weight was measured in lightweight clothes and without shoes, using a scale (Type UM-018 Body Fat & Water Monitor, Tanita Corporation, Tokyo, Japan), to the nearest 0.1 kg. According to these measurements, Body Mass Index (BMI) will be calculated at 13 years through the standard formula $\frac{\text{weight (kg)}}{\text{height}m^2}$, and age and sex-specific BMI standard deviation scores (BMI z-scores) were derived, based on the World Health Organization (WHO) growth reference values [75]. The BMI z-score was categorized as follows: normal weight (> -2 SD and < 1 SD) and overweight/obesity (≥ 1 SD). For simplicity, I will refer to the class overweight/obesity simply as obesity.

5.2 Descriptive analysis

A descriptive analysis of each exposure was performed. For each categorical exposure, the absolute and relative frequency were computed, for each skewed continuous exposure the values of the median, minimum, and maximum were computed, and for each approximately symmetric continuous exposure, the values of the mean and standard deviation were displayed. Sex differences were tested but were not found, so the descriptive analysis was only performed for the total dataset. In order to gain a better understanding of the relationship between each exposure and the outcome variable, a comprehensive descriptive analysis was conducted on each level of the outcome (normal at 13 y.o. or obese

at 13 y.o.). Tables A.1 and A.2 in Appendix A show the results. The majority of participating mothers had primary education, were either married or cohabiting, held paid jobs, and had an average age of 30 years. The household income was generally high among the participants. Moreover, most mothers had a normal weight gain during pregnancy, did not smoke (81%), and did not encounter hypertensive complications (90%) or gestational diabetes (93%). Concerning their children, boys constituted a slightly higher proportion (51%) than girls (49%) and the prevalence of overweight/obesity was 32%, 39%, and 43% at 4, 7, and 10 y.o. respectively. Regarding the outcome, the prevalence of obesity at 13 y.o. was 35%.

Chapter 6

Results - Missing Values

In this chapter, we perform a detailed descriptive analysis of the missing values in the dataset of our case study, and present and discuss, the results from the imputation model.

The longitudinal dataset was divided into four static datasets corresponding to each time point considered in the study (pregnancy, 4, 7, and 10 y.o.). Table 6.1 shows the percentage of missing values and the number of complete cases (observations without missing values) per dataset.

TABLE 6.1: Missing values (number (percentage)) and the number of complete cases, per dataset.

Dataset	Missings	Number of Complete Cases
Pregnancy/Infancy	5828 (5.72 %)	1581
4 Years	4592 (8.3 %)	2747
7 Years	3953 (6.66 %)	3314
10 Years	2238 (5.27 %)	3480
Total Dataset	16290 (7.0 %)	1073

In figures 6.1 and 6.2 we can observe the percentage of missing values per variable in each dataset. For the dataset concerning pregnancy, the maximum percentage of missing values is 24.2%, a moderate level of missing data, and occurs in the BMI of the child at two years old. In the 4 y.o. dataset, the variable active play has the maximum number of missing values, corresponding to a percentage of 17.9%. In the case of the 7 y.o. dataset, the maximum percentage of missing values occurs for the variable maternal BMI, corresponding to a percentage of 12.5%. Finally, the 10 y.o. dataset has the lowest maximum percentage of missing values of 10%, which occurs in the variable maternal education. Overall, the percentage of missing values for each variable is relatively low. However, it

is important to note that the percentage of missing values varies significantly across different variables, which suggests that the missing mechanism in each dataset might not be Missing Completely at Random (MCAR). In fact, in an MCAR scenario, all variables would exhibit a similar proportion of missing values, which is not observed in this case.

Figure 6.3 plots the number of missing values according to the outcome class (normal or overweight/obese). In each dataset, the number of missing values is more pronounced within the normal class. This results from having more observations within the normal class compared to the overweight/obese class. The figure also shows that for the pregnancy, 4 y.o. and 7 y.o. datasets, in general, the percentage of missing values for the obese class is generally slightly greater than one-third of the overall percentage of missing values. This is in accordance with to the prevalence of obesity, which is 35%. On the contrary, for the 10 y.o. dataset, we observe that the variables *sedentary time*, and *sports activity* exhibit a greater percentage of missing values for the obese category than the expected value of 35%. This suggests that the missing values of these variables depend on the outcome class. Typically, this occurs in MAR datasets in which the missing values of one variable depend on the observed values of another variable.

Figures 6.5, 6.6, 6.7 show the location of the observations that contain a missing value, for each variable. In figures 6.6 and 6.7, we see that there are several observations that have a missing value in most of the variables. For a better analysis, for each dataset, the histogram of the frequency of the number of missing values per observation was plotted. Figure 6.8 shows that there are some observations with a lot of missing values so the team decided to exclude the observations that had a missing value in more than half of the variables. In the pregnancy dataset, no observation was excluded. For the other datasets, 116, 310, and 120 observations were excluded for the 4-year, 7-year, and 10-year datasets respectively.

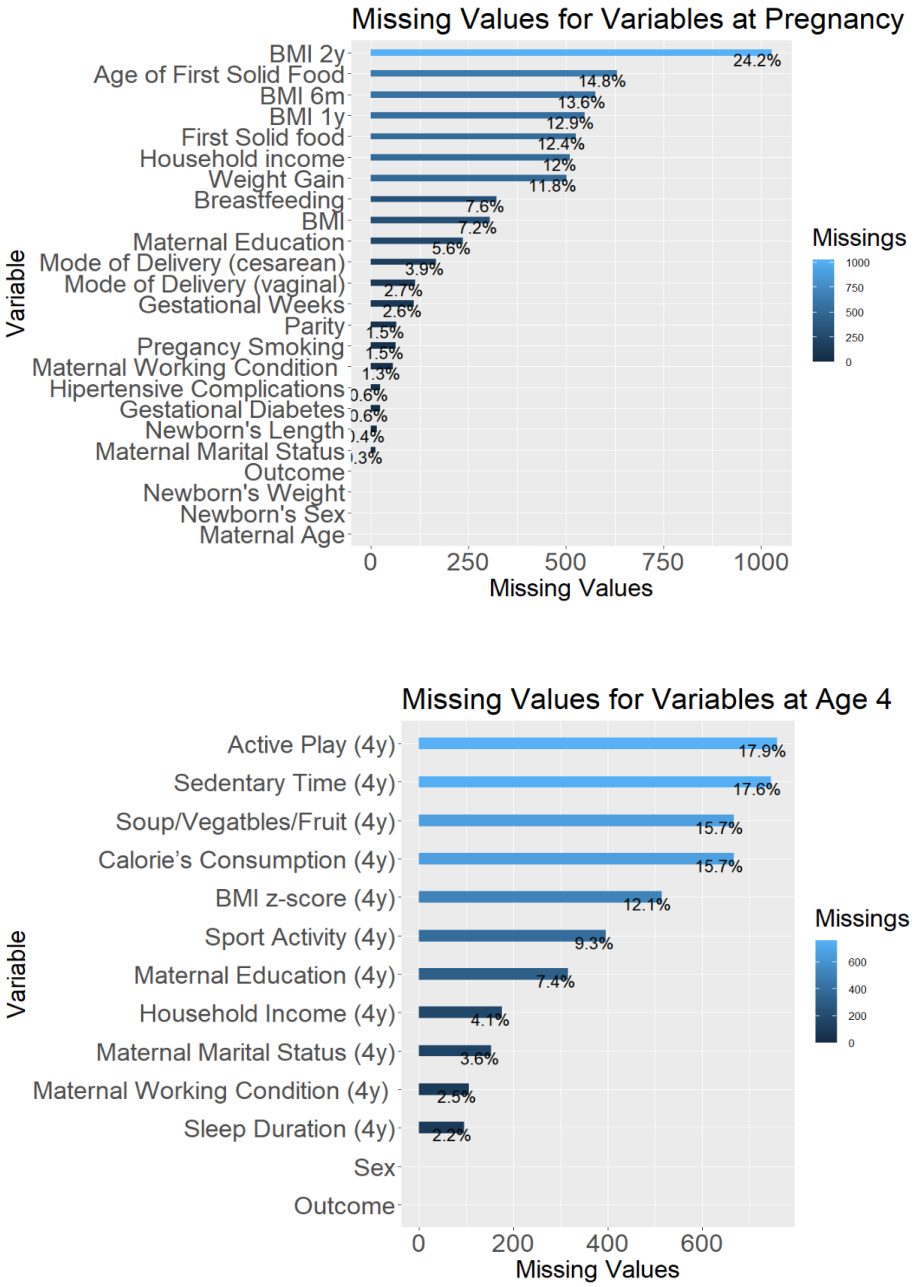


FIGURE 6.1: Missing values per variable for the pregnancy and 4 y.o. dataset.

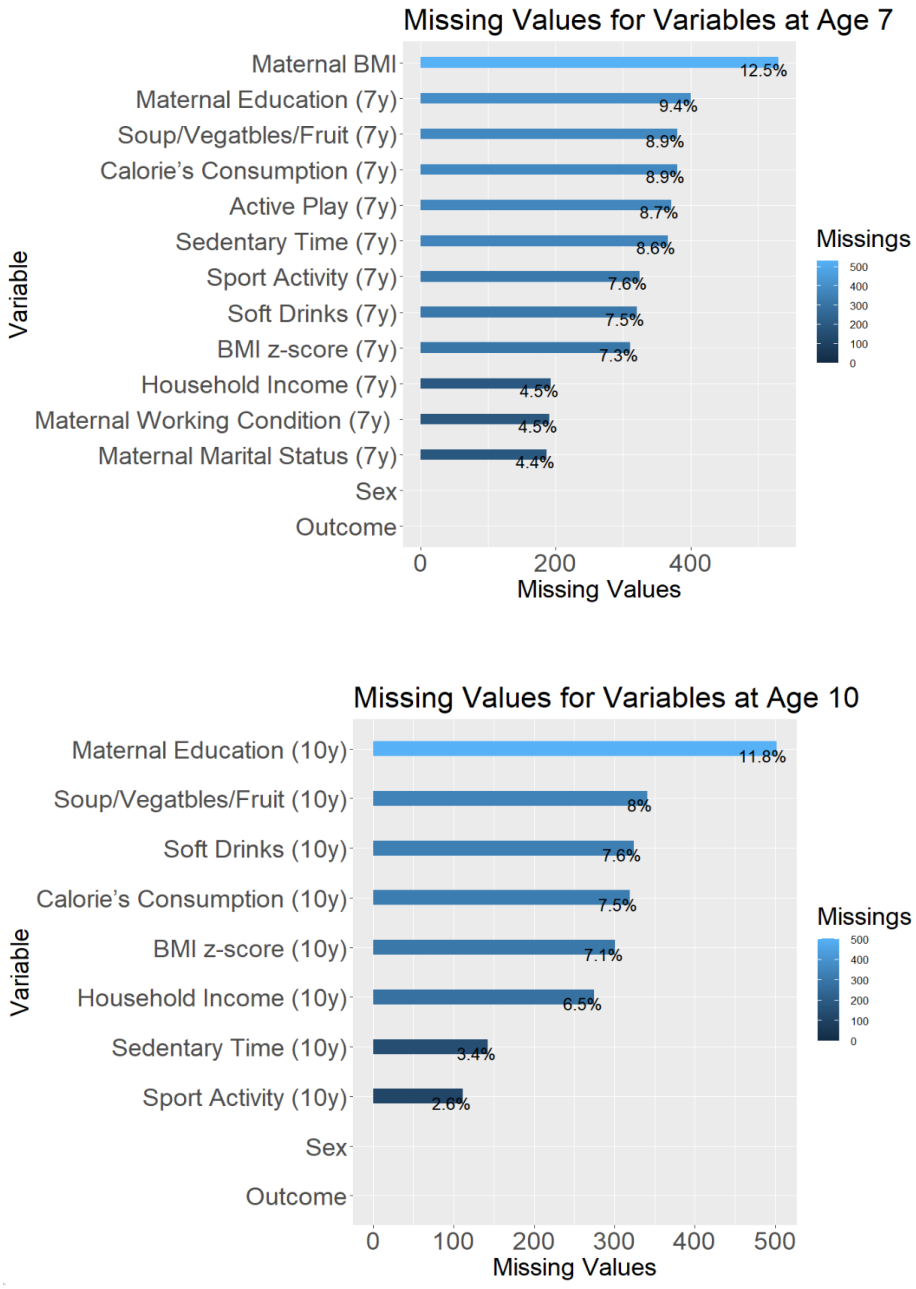
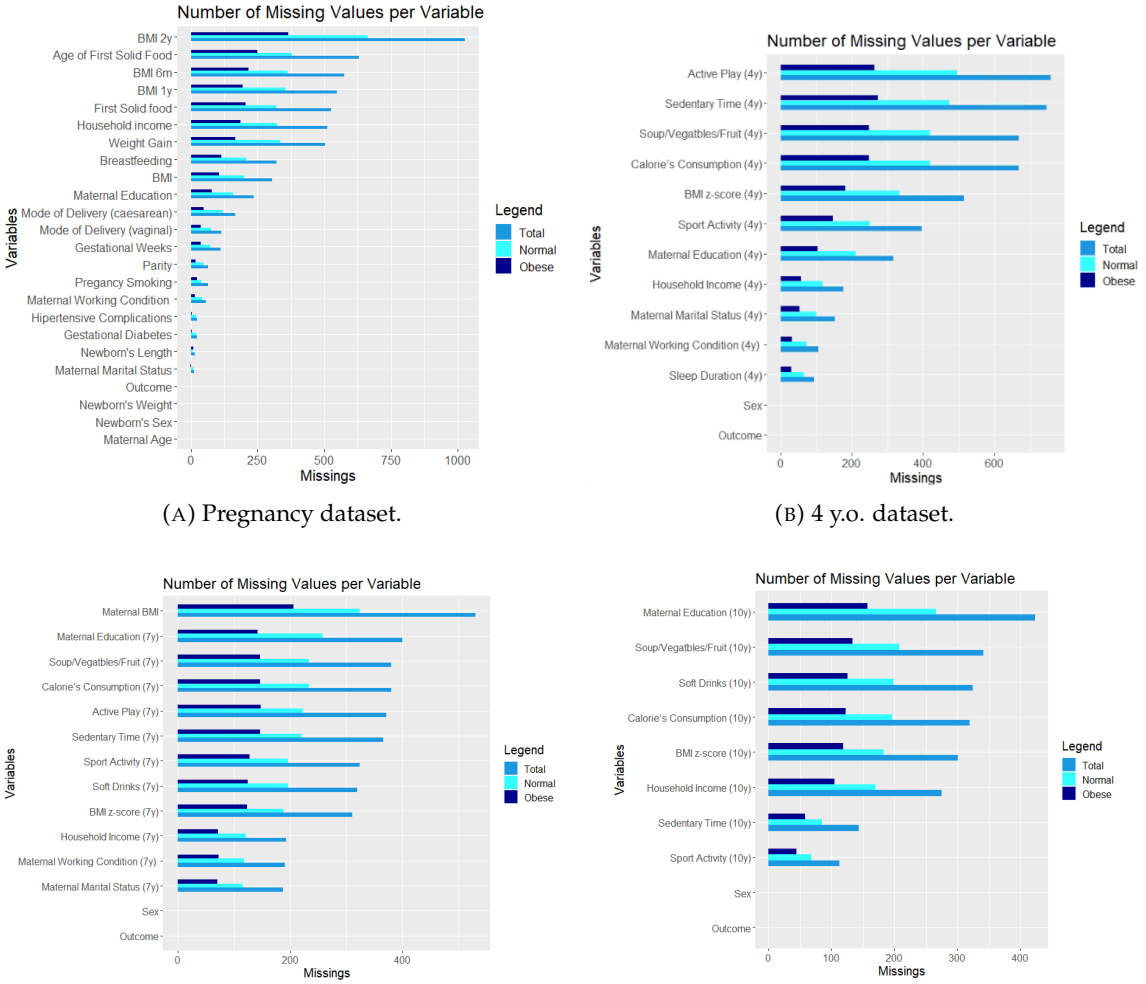


FIGURE 6.2: Missing values per variable for the 7 and 10 y.o. datasets.



(A) Pregnancy dataset. (B) 4 y.o. dataset.

(C) 7 y.o. dataset. (D) 10 y.o. dataset.

FIGURE 6.3: Missing values per variable for each time-point dataset according to the outcome class.

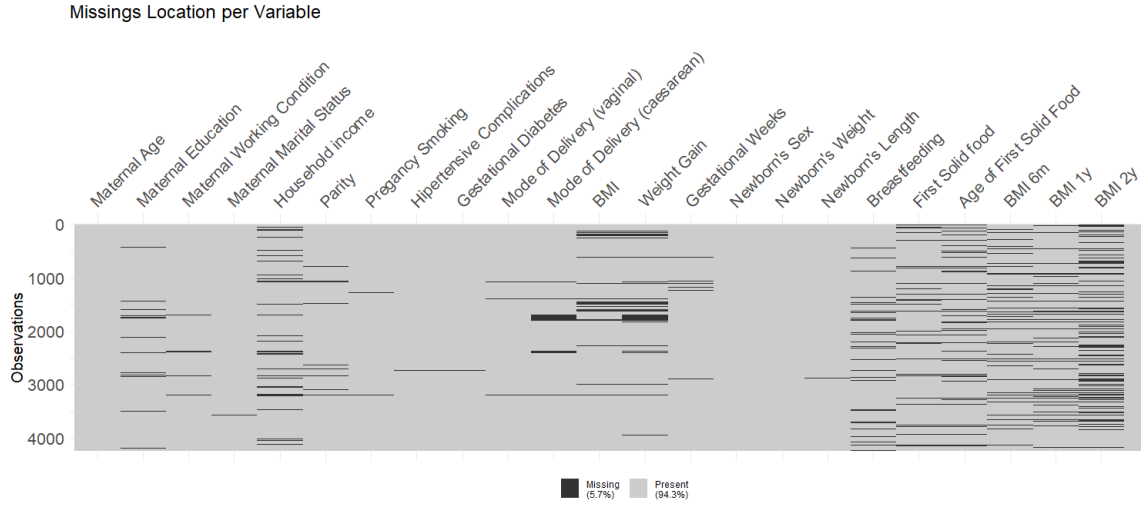


FIGURE 6.4: Location of the missing values per variable for the pregnancy dataset.

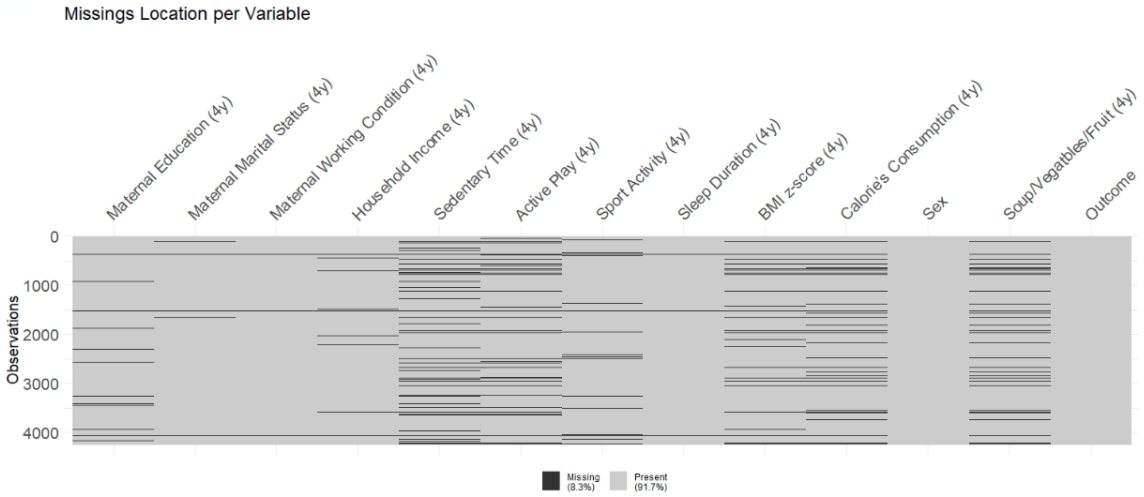


FIGURE 6.5: Location of the missing values per variable for the 4 y.o. dataset.

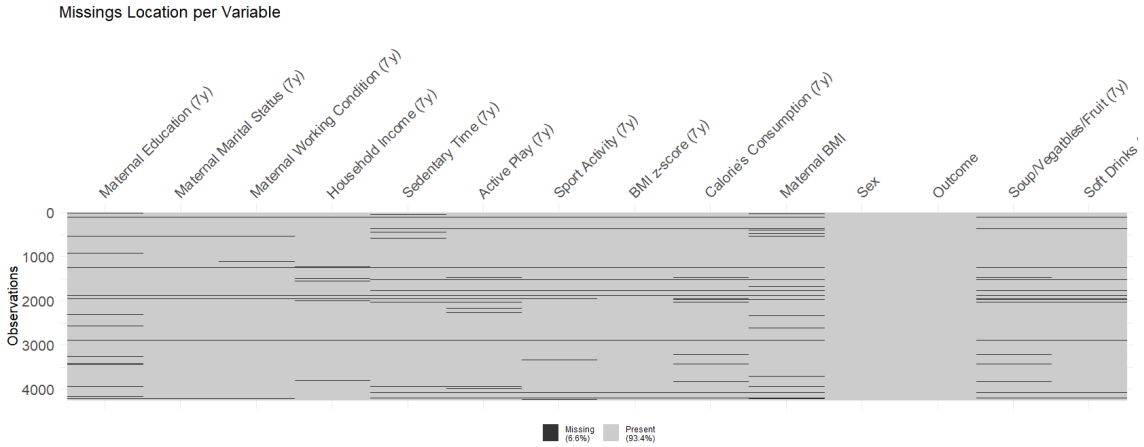


FIGURE 6.6: Location of the missing values per variable for the 7 y.o. dataset.

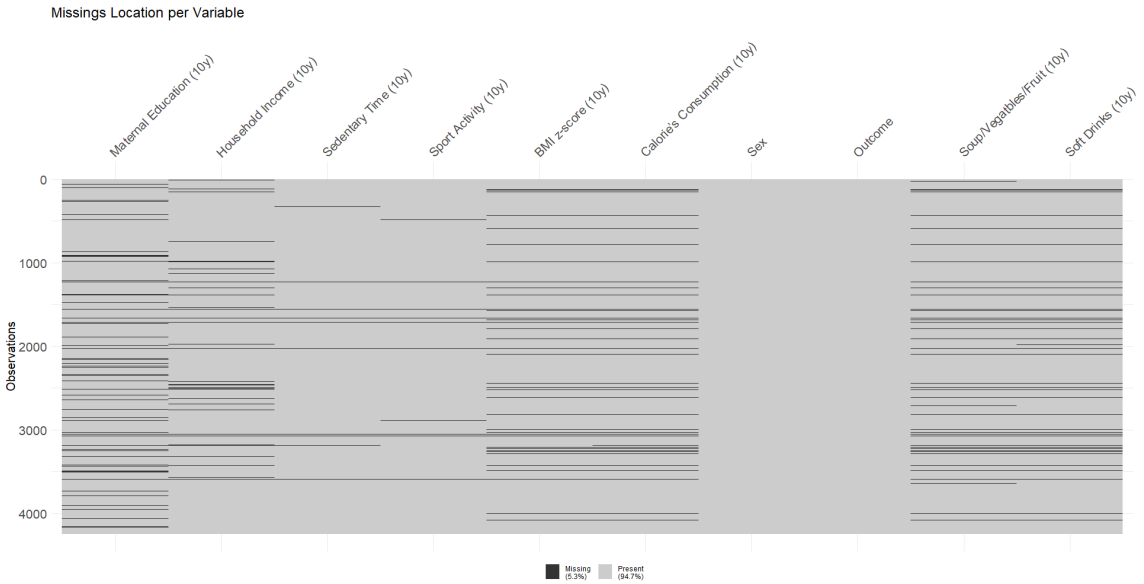


FIGURE 6.7: Location of the missing values per variable for the 10 y.o. dataset.

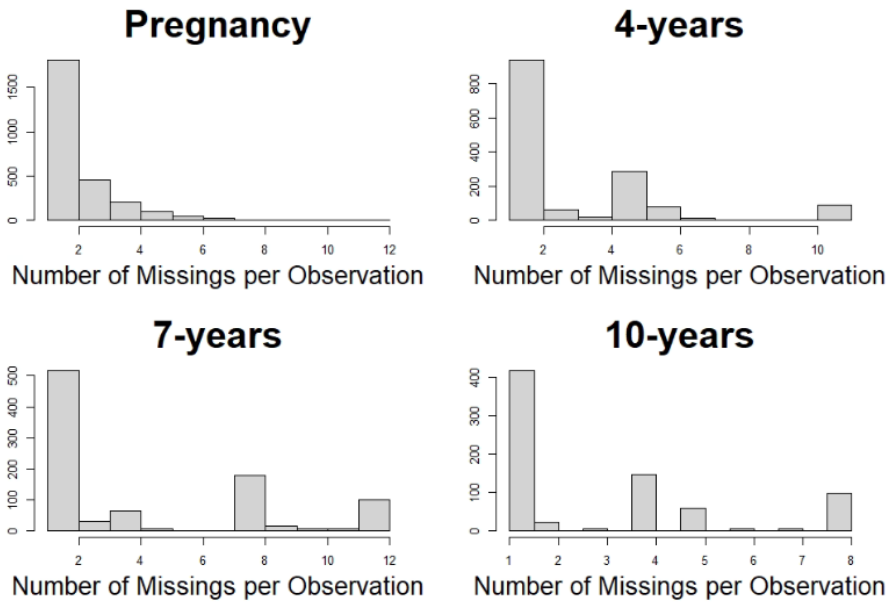


FIGURE 6.8: Histograms for the number of missing values per observation, per dataset.

6.1 Missing Values Analysis

The following subsections present a simulation study performed on the pregnancy dataset. The analysis was only conducted on this dataset since it was the first dataset that was made available to us. This analysis was performed with the objective of understanding three concerns regarding missing values: (1) which imputation method is the most appropriate for the dataset under analysis; (2) how similar are the imputed values across the multiply imputed datasets; (3) which imputation method obtains imputed values more similar to the original observed value.

6.1.1 Comparison of Multiple Imputation Methods

It was necessary to choose which imputation technique should be implemented. With this in mind, a simulation study on a subset of the dataset was conducted. This subset corresponds to the following variables regarding the pregnancy stage and the newborn: mother's age, working status, marital status, years of education, income, number of previous pregnancies, smoking habits during pregnancy, gestational hypertension and diabetes, BMI, weight gain, number of gestational weeks, number of twins in the current pregnancy, sex, length, weight at birth, outcome. It contained 4237 observations, 17 variables, and 5.38% of missing values. This dataset is not equal to the pregnancy dataset presented earlier because later on in the internship, other variables were included.

Our first main goal was to do a static model for predicting childhood obesity at age 13 considering only the pregnancy variables. Since the outcome is a binary variable, logistic binary regression was the selected method. We inspected the effect of six different imputation methods on the results from a logistic regression model in order to compare them and select the best one - Mean/ Mode imputation, Random imputation, Hot Deck imputation, and MICE with 3 imputation models. The first imputation model (MICE-Pmm) included predictive mean matching to impute continuous variables, binary logistic regression to impute binary variables, and multinomial logistic regression to impute categorical variables with more than 2 levels. The second model (MICE-Lm) considered a linear model to impute continuous variables, binary logistic regression to impute binary variables, and multinomial logistic regression to impute categorical variables with more than 2 levels. Finally, the third model consisted of a random forest with classification trees for the categorical variables, and regression trees for the continuous variables. The first imputation model was chosen because it is the default method in R when using MICE,

the second model was chosen because it is a parametric model and linear models are one of the most common regression methods, and the third model was considered because it is a non-parametric model.

The simulation study was carried out in the following way. Firstly, a subset (n=3504) of the complete cases of the entire dataset (n=4237) was considered, and a logistic regression model predicting obesity was performed. This is equivalent to performing listwise deletion since we are deleting the observations with missing values, and fitting the model in the remaining observations. In this complete version of the dataset, a proportion of missing values equal to the proportion of missing values in the initial dataset (5.38%) was randomly generated, giving rise to a dataset with an MCAR mechanism. This was performed using the package *missMethods*, in R. The process was repeated 150 times. The six imputation methods were performed for every new dataset with missing values, followed by the logistic regression model predicting the outcome variable. For each imputation method and generated dataset, the model's coefficients, respective standard deviations, and corresponding p-values were retained. Since the simulation was repeated 150 times, each parameter in each imputation method has 150 values. Finally, to compare the results from all the methods with the baseline model (applied to the dataset with complete cases), histograms with these 150 values for every parameter were plotted. Since there are $(26 \times 3) \times 6$=468 histograms, only the worst and best scenarios of each model are presented here. In addition, the red point in each histogram corresponds to the parameter estimated by the model fitted to the complete dataset.

Figure 6.9 provides relatively symmetric distributions for the coefficient values. The histograms for the p-values are right-skewed. We can also note that some of the displayed histograms lack the red point, meaning that the coefficient from the complete dataset was "far" from the actual value. This was prevalent in the estimation of the standard error of several coefficients. The imputation method MICE-Pmm was the only method for which all histograms contained the red point. Despite this, the histogram analysis is not enlightening enough to decide on the best imputation model for this dataset. Therefore, the order of the sample quantile of the actual value (redpoint) was computed for every parameter in every model considering the sample of 150 values of every parameter.

To evaluate the overall results, the mean, standard deviation, median, minimum, and maximum for the quantiles orders are displayed in table 6.2 for every model according to the estimates (coefficients, p-values, standard deviations).

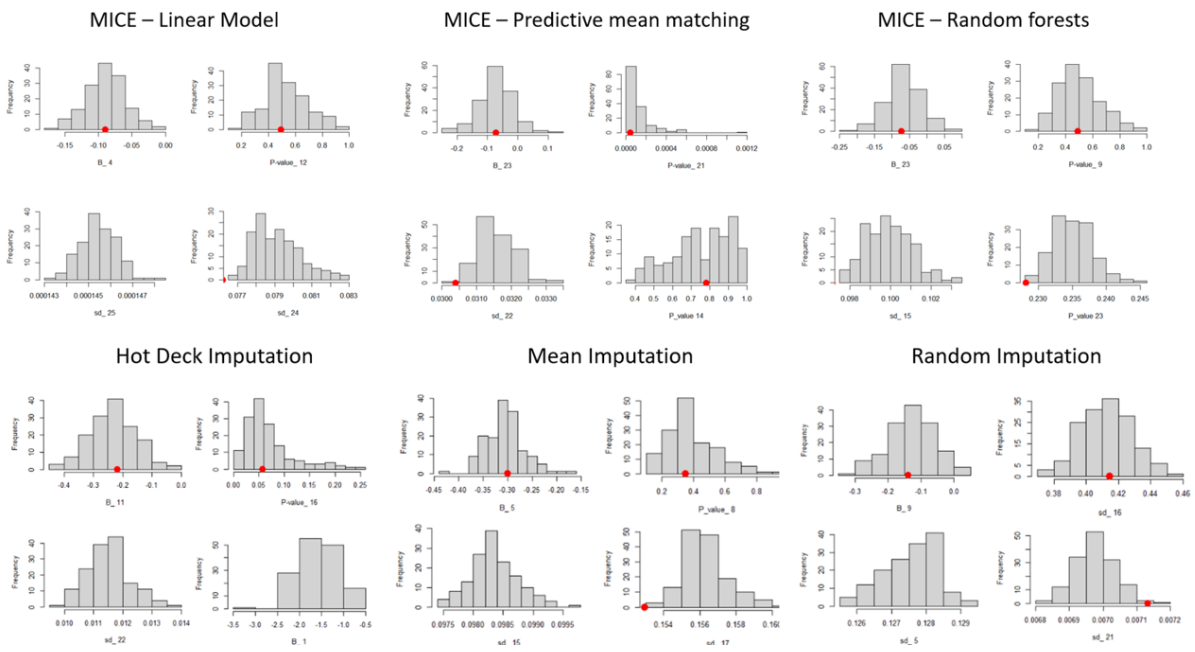


FIGURE 6.9: Histograms for the best (upper row) and worst (lower row) situations for the six imputation models.

The values presented in table 6.2 indicate that MICE-Pmm was the most consistent method across the values of the coefficients since the mean was 0.49 and the values had a range of 0.19 (maximum-minimum). This means that for all coefficients, the order of the quantile was near 0.5 which is desired. Although the other methods returned a mean/-median close to 0.5, the standard deviation, minimum, and maximum showed that there are several coefficients in which the order of the quantile was far from 0.5. The values for the p-value estimates were similar across all methods except for the MICE-Rf, in which mean and median values were far from 0.5. Therefore, the MICE-Pmm appears to be the best imputation model for the dataset under analysis. The values for the standard error were unsatisfactory for all imputation models. In order to ensure that there was actually a difference between the six methods, an ANOVA model was conducted, followed by multiple comparisons using the Tukey correction. The results are shown in figure 6.10. For the coefficients, there was no significant difference between the methods. For the p-values, there was a significant difference between method 3 (MICE-Rf) and 4 (Hot Deck) which is suggested by looking at table 6.2 since the average for MICE-Rf was 0.36 and for Hot-Deck was 0.52. For the standard errors, almost every method differed from one another, meaning that the standard error estimates were very unstable.

Summing up, the previous simulation study suggested that multiple imputation with chained equations using predictive mean matching, logistic regression, and polytomous

TABLE 6.2: Summary of the quantiles orders for each model; the best values are in blue.

	MICE-Lm	MICE-Pmm	MICE-Rf	Hot Deck	Mean	Random
Coefficients						
average	0.47	0.49	0.52	0.49	0.51	0.55
median	0.48	0.49	0.53	0.48	0.49	0.53
sd	0.11	0.05	0.14	0.21	0.16	0.24
min	0.23	0.40	0.23	0.00	0.13	0.06
max	0.78	0.59	0.79	1.00	0.84	0.98
P-values						
average	0.44	0.40	0.36	0.52	0.45	0.42
median	0.42	0.40	0.38	0.52	0.45	0.42
sd	0.16	0.15	0.19	0.18	0.19	0.21
min	0.04	0.02	0.01	0.13	0.04	0.03
max	0.82	0.85	0.89	1.00	0.90	0.97
Standard Errors						
average	0.02	0.02	0.04	0.92	0.33	0.73
median	0.00	0.00	0.02	0.99	0.073	0.70
sd	0.05	0.04	0.05	0.23	0.41	0.13
min	0.00	0.00	0.00	0.59	0.00	0.35
max	0.15	0.13	0.16	1.00	1.00	1.00

regression as the imputation models for continuous, binary, and categorical variables, respectively, was the method with the best results for the dataset under analysis. These first results were presented at the scientific meeting JOCLAD 2023, having actually deserved a conference grant for attendance at the meeting and oral presentation of the work.

Finally, a comparison between the 3 imputation models used in MICE was performed. The values for the relative efficiency (RE), the fraction of missing information (FMI/ γ), and the relative increase in variance (RIV) were computed. As the simulation was repeated 150 times, every variable provided 150 values for each of RE, RIV, and γ. Histograms for each variable, each measure, and each model were plotted. The graphs were very similar across variables hence figure 6.11 only shows 4 histograms for each measure and each imputation model. The histograms are very similar for the 3 imputation methods. Moreover, the distributions of RIV and FMI suggest a right-skewed distribution, while the distribution of RE suggests a left-skewed distribution.

For a better analysis of the results, table 6.3 shows the median, minimum and maximum value, for each measure, and for each imputation model (for each variable, the median of 150 values was computed).

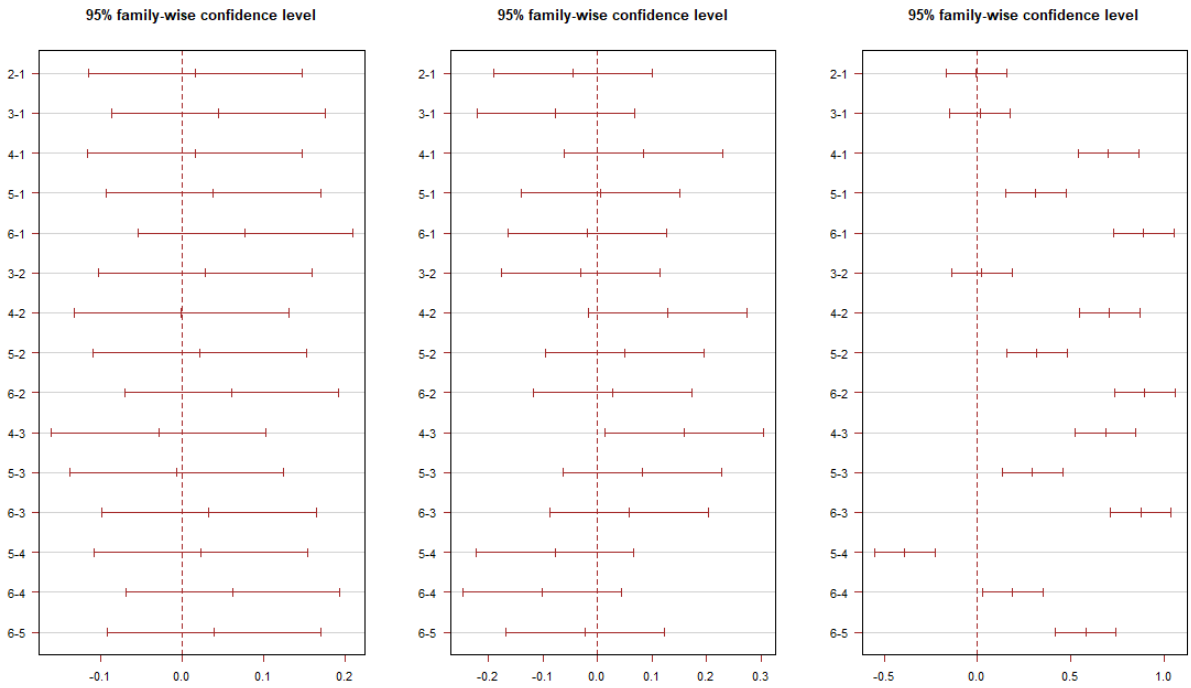


FIGURE 6.10: Confidence intervals by the multiple comparisons procedure (left: coefficients; middle: p-values; right: standard errors).

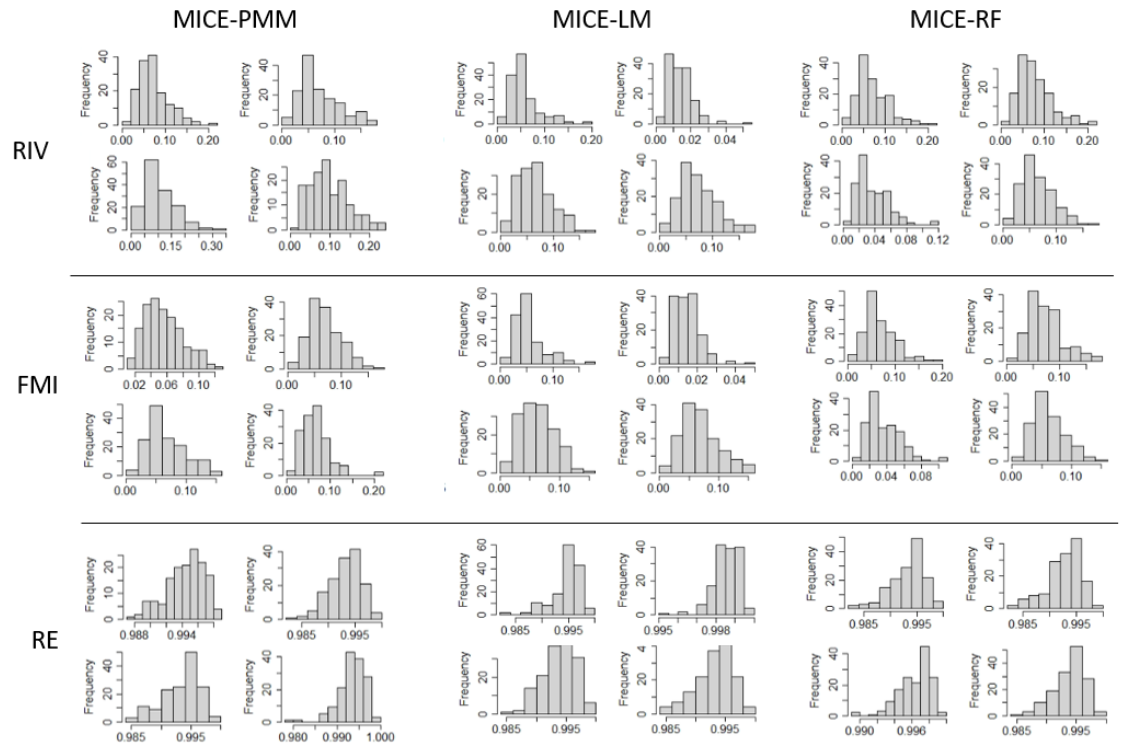


FIGURE 6.11: Histograms for RIV, FMI, and RE, for each imputation model.

TABLE 6.3: Median (minimum-maximum) of the relative increase in variance (RIV), the fraction of missing information (FMI), and relative efficiency (RE), for each model.

	RIV	FMI	RE
Predictive Mean Matching	0.067 (0.053-0.106)	0.064 (0.051-0.098)	0.994 (0.990-0.995)
Linear Model	0.064 (0.007-0.087)	0.061 (0.007-0.082)	0.994 (0.991-0.999)
Random Forest	0.063 (0.023-0.135)	0.060 (0.024-0.122)	0.994 (0.988-0.998)

Looking at the values RIV, FMI, and RE presented in table 6.3 we see that they are very similar for the 3 imputation methods and therefore conclude that the relative efficiency of the imputation does not vary much with the imputation model.

6.1.2 Concordance Measures between Multiple Imputed Values

In this subsection, we investigate the variability of the imputed values across the m imputed datasets and their relation with the corresponding imputation model. A simulation study similar to the one presented in 6.1.1 was performed. The study was conducted on a subset of the entire dataset containing the variables: mother's age, working status, marital status, years of education, income, parity, smoking habits during pregnancy, gestational hypertension and diabetes, vaginal as the mode of delivery, cesarean as the mode of delivery, BMI, weight gain, number of gestational weeks, sex, length, weight at birth, outcome. The subset contains 18 variables, 3139 observations, and 3.6% missing values. This subset was slightly different from the dataset used before because the research team at the internship decided to exclude and include some different variables. Since this dataset contained 3.6% of missing values, the same percentage (3.6%) of missing values were generated on the complete dataset, again according to an MCAR mechanism. This resulted in 112 generated missing values per variable. Afterward, MICE with each of the 3 different imputation models presented in 6.1.1 was applied. The concordance between the imputed values generated by each of the methods was evaluated by Fleiss' kappa for categorical variables and the Intra-Class Correlation Coefficient (ICC) for continuous variables. Since the multiple imputation procedure created 10 imputed datasets, each variable had 10 different versions. The concordance measure was computed using these 10 different versions but using only the observations from that variable that were missing before applying the imputation method.

TABLE 6.4: Summary of the agreement measurements for each variable and each imputation method.

Method	ICC			Fleiss' kappa					
	Age	G. Weeks	BMI	Inc.	Educ	Work	Marital	Parity	Smoking
Linear	0.982	0.996	0.984	0.239	0.218	0.174	0.062	0.280	0.059
Model	N. Weight	N. Weight		Hip.	Diab.	Vaginal	Ceasearen	Weight G.	Sex
	0.999	0.999		0.611	0.668	0.930	0.923	0.099	0.027
	Age	G. Weeks	BMI	Inc.	Educ	Work	Marital	Parity	Smoking
Random	0.237	0.438	0.110	0.245	0.099	0.117	0.291	0.256	0.014
Forest	N. Weight	N. Weight		Hip.	Diab.	Vaginal	Ceasearen	Weight G.	Sex
	0.488	0.406		0.692	0.532	0.983	0.978	0.100	0.078
	Age	G. Weeks	BMI	Inc.	Educ	Work	Marital	Parity	Smoking
Predictive	0.298	0.579	0.227	0.232	0.134	0.185	0.231	0.292	0.036
Mean	N. Weight	N. Weight		Hip.	Diab.	Vaginal	Ceasearen	Weight G.	Sex
Matching	0.701	0.659		0.598	0.695	0.942	0.955	0.076	0.014

Table 6.4 shows the results for the best, worst, and median values of the agreement measures, for each variable and each imputation model. For the continuous variables, the best results were obtained for the linear model, displaying very high values. For the categorical variables, the concordance measures were lower and it is not possible to identify the imputation model with the best results since the values depend on the variable under analysis.

6.1.3 Imputed Values vs Observed Values

The imputed values were afterward compared with the originally collected values (since we generated the missing values on the complete dataset, we have the original observed value that later on was set to missing). For each categorical variable, we counted the proportion of imputed classes matching exactly the original class. For each continuous variable, we standardized each variable and computed the bias, mean squared error (MSE), and mean absolute percentage error (MAPE) [76].

The results for the proportion of matching for the categorical variables were the following:

Predictive Mean Matching- 71 %; **Linear Model-** 73 % ; **Random Forest-** 76 %.

showing that the proportion of matching among the imputed values and the original values of the categorical variables was higher for the random forest imputation model.

The calculation of the bias, mean squared error (MSE) and mean absolute percentage error (MAPE) of the estimates for the continuous variables followed the following procedure. We fixed one of the observed values on the original complete dataset that was later

on set to missing when the missing values were generated. After the multiple imputation procedures, since m was equal to 10, this observed value had 10 different imputed values that were trying to estimate it. To compute the bias, MSE, and MAPE of the estimates, the observed value was considered as the true parameter, and the vector of the 10 imputed values served as estimates. This procedure was repeated for each originally observed value that was subsequently set to missing. The histograms of the performance measures for each variable and each imputation model are presented in figures A.1, A.2, A.3 in the Appendix A. Most histograms exhibit a skewed distribution, particularly for MAPE and MSE values. Table 6.5 shows the median value for each variable, performance measure, and imputation model.

TABLE 6.5: Median values for the performance measures, for each variable and each imputation model.

	Bias			**MSE**			**MAPE**		
Variable	**Pmm**	**Lm**	**Rf**	**Pmm**	**Lm**	**Rf**	**Pmm**	**Lm**	**Rf**
Mother's Age	0.13	0.10	0.05	1.07	0.40	1.16	1.44	0.93	1.44
Mother's BMI	-0.04	-0.01	0.01	1.06	0.25	0.93	1.56	0.90	1.47
Gestational weeks	-0.08	-0.12	-0.08	0.71	0.26	0.77	1.25	0.98	1.19
Newborn's Weight	-0.02	0.02	0.05	0.40	0.13	0.62	0.92	0.57	1.05
Newborn's Length	0.03	-0.14	-0.12	0.52	0.15	0.70	0.93	0.62	1.05

The table indicates that the linear model performed better for MSE and MAPE. Only looking at the results obtained for the bias, it is difficult to identify the imputation model that is providing imputed values more similar to the actually observed values.

6.2 Imputation

The imputation procedure was carried out separately for each time-point dataset to avoid multicollinearity problems. The method of Multiple Imputation with Chained Equations (MICE) was chosen to handle missing values. The imputation model used predictive mean matching to impute continuous variables, binary logistic regression to impute binary variables, and multinomial logistic regression to impute categorical variables with more than 2 levels. MICE was performed resulting in 10 imputed datasets. To evaluate the impact of the imputation procedure on the variables, tables A.3, A.4, A.5, and A.6 in Appendix A show the descriptive analysis for all variables before and after imputation.

Looking at the tables, it is possible to conclude that the values remained unchanged. Consequently, the imputation process did not relevantly alter the overall distribution of the variables in each dataset.

Chapter 7

Results - Models

Firstly, and in order to examine the association of each variable separately with the response (crude effects), separate logistic regression models were fitted for each variable. However, since this is a scenario of multiple comparisons, the Holm procedure was performed to control the Family-Wise Error Rate (FWER). The procedure works by sorting out the p-values obtained from the multiple tests in ascending order. Then, each p-value is compared to a series of adjusted significance levels, starting from the most significant one. If a p-value is smaller than or equal to the adjusted significance level, the corresponding null hypothesis is rejected. If a p-value is larger than the adjusted significance level, the remaining p-values are not considered any further, and their associated null hypotheses are not rejected [77]. The adjusted significance level at step j is $\frac{\alpha}{m-j+1}$, where m is the total number of tests and α is the defined value for the FWER.

After examining the crude effects, the models were fitted considering all exposures. The results are presented separately, for the static and the dynamic models. All models are presented without interactions since these terms did not improve the predictive performances.

For the evaluation of the model's performance, each dataset was divided into a train set (70 %) and a test set (30 %). All models were fitted on the train set and the performance measures were computed on the test set.

7.1 Static Models

7.1.1 Pregnancy/Infancy Model

The first static model for obesity at the age of 13 includes all exposures regarding the pregnancy stage and the infancy of the child, summing up 23 exposures. Initially, all crude effects were inspected. Table A.7 in Appendix A shows the results for the corresponding *odds ratio* and *adjusted p-values*. The exposures maternal education, working condition, household income, smoking habits during pregnancy, hypertensive complications, gestational diabetes, mother's pre-conception BMI, weight gain during pregnancy, first solid food introduced, and child's BMI at 6 months, 1 year, and 2 years are significantly associated with the child being overweight/obese at age 13. Figure 7.1 shows a visual representation of the adjusted p-values. It is possible to see that the BMI of the child and of the mother have the most significant association with the outcome.

For the pregnancy/infancy model including all exposures, logistic regression could not be used due to the multicollinearity caused by using the child's BMI at 6 months, 1 year, and 2 years. In fact, the Pearson correlation between the child's BMI at 6 months and 1 year is estimated at 0.72. A scatterplot of the variables is presented in Figure 7.2.

To incorporate the three BMI longitudinal measures, two approaches were considered. Firstly, we conducted linear regression models for each child, treating each BMI measure as the dependent variable and the time of measurement as the independent variable. For each subject, we saved the values for the intercept and the slope. Thus, the first approach was to perform logistic regression with the pregnancy/infancy exposures and the intercept and slope for each subject instead of using each of the three BMIs. The second approach consisted of applying a penalized regression method. Elastic Net (ENET) was the chosen method since it includes both the L1 penalty (similar to LASSO) and the L2 penalty (similar to Ridge regression) in the objective function. Moreover, this method can handle situations with a large number of correlated predictors and perform both feature selection and regularization. Since we had multiply imputed datasets, the method described in section 3.4 with a stacked objective function was followed.

Results from the Logistic Model

Before fitting the model, the relationship between the continuous predictors and the logit of the response was assessed to ensure the presence of a linear relationship. Figure 7.3 suggests that the linearity assumption holds.

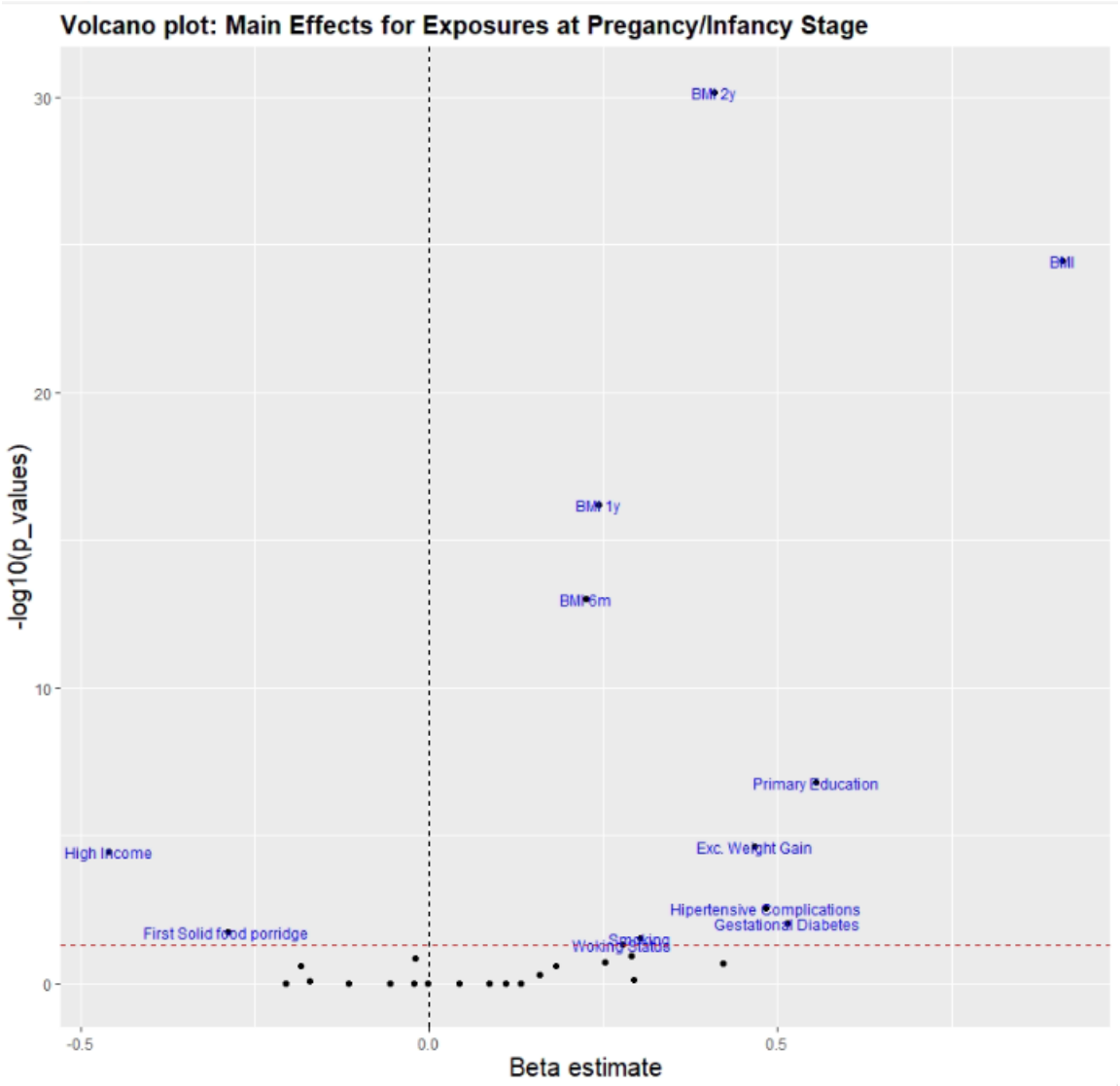


FIGURE 7.1: Associations between pregnancy/infancy exposures and the BMI z-score at age 13. The volcano plot shows the p-values against the beta coefficient. Red dashed horizontal line at the value of FWER=0.05.

The final model was obtained by backward elimination. The results were combined by Rubin's Rule and are shown in table 7.1. We conclude that:

1. For each year of age of the mother, the odds of the child being obese at age 13 decreases by 2%.

2. There are no significant differences between the effects of low household income and middle household income. However, children from households with high income have approximately 23% lower odds of being obese at age 13 than children from low-income households.

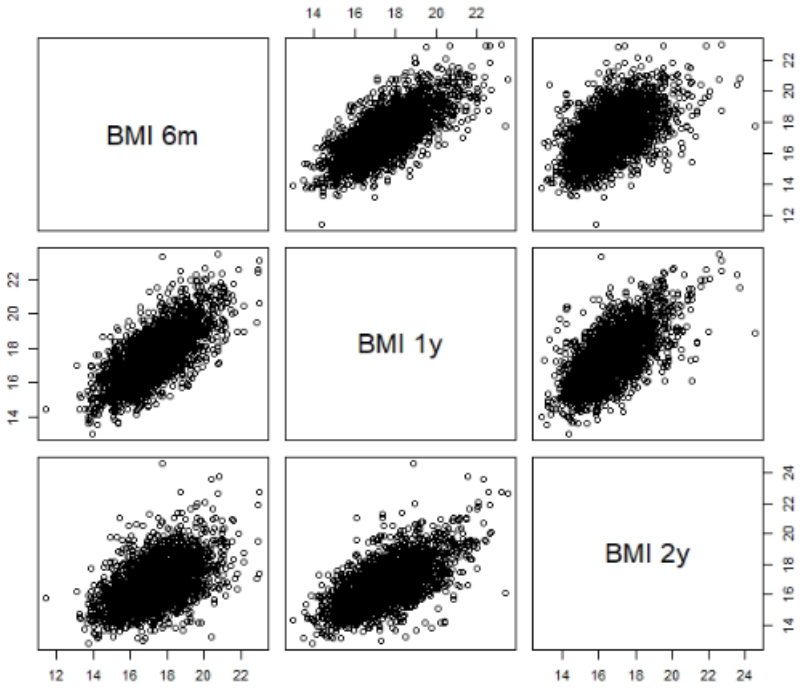


FIGURE 7.2: Scatter plot of Body Mass Index (BMI) at 6 months, 1-year-old, and 2 years old.

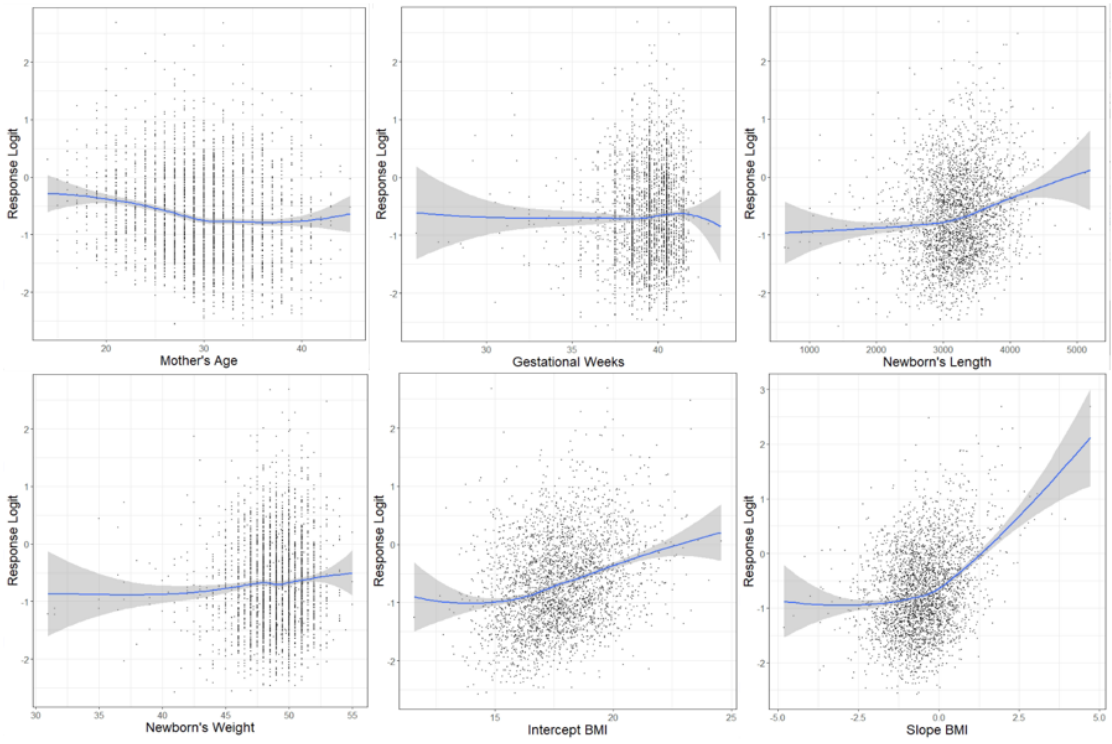


FIGURE 7.3: Scatter plot of the logit of the response and each continuous variable, for the pregnancy dataset. The regression line (in blue) was obtained by loess, and its confidence interval is pictured in grey.

TABLE 7.1: Odds Ratios and p-values for every variable present in the final model.

Exposures	Odds-Ratio (95 % CI)	p-Value
Age	0.98 (0.96, 0.99)	0.008
Household Income		
Low	Ref	Ref
Middle	0.87 (0.70, 1.07)	0.19
High	0.77 (0.61, 0.96)	0.02
Smoking Habits during Pregnancy		
Never smoked	Ref	Ref
Smoked	1.40 (1.14, 1.72)	0.001
Hypertensive complications		
Yes	1.64 (1.25, 2.14)	< 0.001
No	Ref	Ref
Pre-conceptual BMI		
Underweight/normal	Ref	Ref
Overweight/obese	2.35 (1.98, 2.80)	< 0.001
First solid food		
Cereal porridge	0.84 (0.70, 0.99)	0.04
Fruit	0.79 (0.52, 1.19)	0.27
Soup	Ref	Ref
Intercept of the BMI Regression	1.48 (1.38, 1.58)	< 0.001
Slope of the BMI Regression	2.02 (1.74, 2.35)	< 0.001

3. Children whose mothers smoked during pregnancy have approximately 40% higher odds of being obese at age 13 than children whose mothers have never smoked.

4. Children whose mothers had hypertensive complications during pregnancy have approximately 64% higher odds of being obese at age 13 than children whose mothers did not have hypertensive complications.

5. Children whose mothers had been overweight/obese before pregnancy have approximately 135% higher odds of being obese at age 13 (is 35 % greater) than children whose mothers had pre-conceptual underweight/normal BMI. This odds ratio indicates that the pre-conceptual BMI of the mother is the most important predictor associated with obesity at age 13. Therefore, prevention should heavily rely on this factor.

6. There are no significant differences between the effects of introducing fruit and introducing soup as the first solid food. Children who were given cereal porridge as their first solid food have approximately 16% lower odds of being obese at age 13 than children who were given soup as their first solid food. However, this interpretation is not entirely correct because pediatricians often advise mothers to introduce

cereal porridge as the first solid food if the baby has a low weight. So actually we cannot interpret the cereal porridge as a factor that decreases the risk of obesity.

7. The (significant) effect of the slopes of the BMI Regressions indicates that, for every unit increase in the rate of change in BMI between two consecutive measurements, the odds of the child being obese at age 13 increases by 102% (is 2 % greater).

Results from the stacked Elastic Net

We used the R function *cv.saenet* from the package *miselect*. The function performs 5-fold cross-validation to select the optimal values for α and λ (parameters of the objective function defined in 3.4). It returned $\alpha = 0.1$ and $\lambda = 0.0036$. Table 7.2 shows the results of the odds-ratio values for the selected variables. The confidence intervals for the odds ratio were not presented because the authors that developed this method did not present how the confidence intervals should be computed. Looking at the table we see that most of the variables (18 out of 23) have been selected, and therefore the low value for α is not highly penalizing the variables. Additionally, the variable with the largest effect on the outcome is the mother's pre-conceptional BMI. The odds ratio shows that children with mothers who were overweight/obese before pregnancy have 18% higher odds of being obese at age 13 than children having mothers with normal weight. Other risk factors for children being obese at the age of 13 are having mothers with low education, that do not engage in paid job activity, lacking a partner, and having a lower household income.

7.1.2 Model for 4 years-old

The 4 y.o.-model includes 11 exposures that were collected at the 4 y.o. of the child. Once again, before fitting the model with all exposures, the crude effects of each exposure were investigated. Table A.8 in Appendix A presents the estimated odds ratios and adjusted p-values. The exposures maternal education, household income, child's sedentary time, and child's BMI are significantly associated with the child being overweight/obese at age 13. Figure 7.4 shows a visual representation of the adjusted p-values. We can see that the BMI of the child had the most significant effect on the outcome.

As the goal of this study was to develop models with good prediction ability, different classification models were fitted in order to select the most suitable. The methods K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Classification Trees were tested. Table A.11 in Appendix A shows the results for those models' performance

TABLE 7.2: Odds ratio for the variables in the final elastic net model.

Exposures	**Odds Ratio**
Maternal Age	0.997
Maternal Education	
Primary	1.013
Secondary	1.00
Tertiary	Ref
Maternal Marital Status	
Married or cohabiting	Ref
No partner	1.065
Working Condition	
Paid job	Ref
No paid job activity	1.008
Household Income	
Low	Ref
Middle	1.000
High	0.974
Smoking During Pregnancy	
No	Ref
Yes	1.054
Hypertensive Complications	
No	Ref
Yes	1.059
Gestational Diabetes	
No	Ref
Yes	1.051
Vaginal Delivery	
Yes	0.987
No	Ref
Maternal BMI	
Underweight/Normal	Ref
Overweight/Obese	1.185
BMI at 6 months	1.012
BMI at 2 years	1.068
Parity	
Nullipara	Ref
Multipara	1.021
Gestational Weight Gain	
Insufficient	1.017
Adequate	Ref
Excessive	1.032
Sex	
Boy	0.987
Girl	Ref
Breastfeeding	
Never	1.014
0-2.9 months	0.982
3-5.9 months	0.986
$\geq$ 6 months	Ref
First Solid Food	
Cereal porridge	0.972
Fruit	0.978
Soup	Ref
Age of First Solid Food	
$<$4 months	1.016
4-5 months	Ref
$\geq$ 6 months	1.005

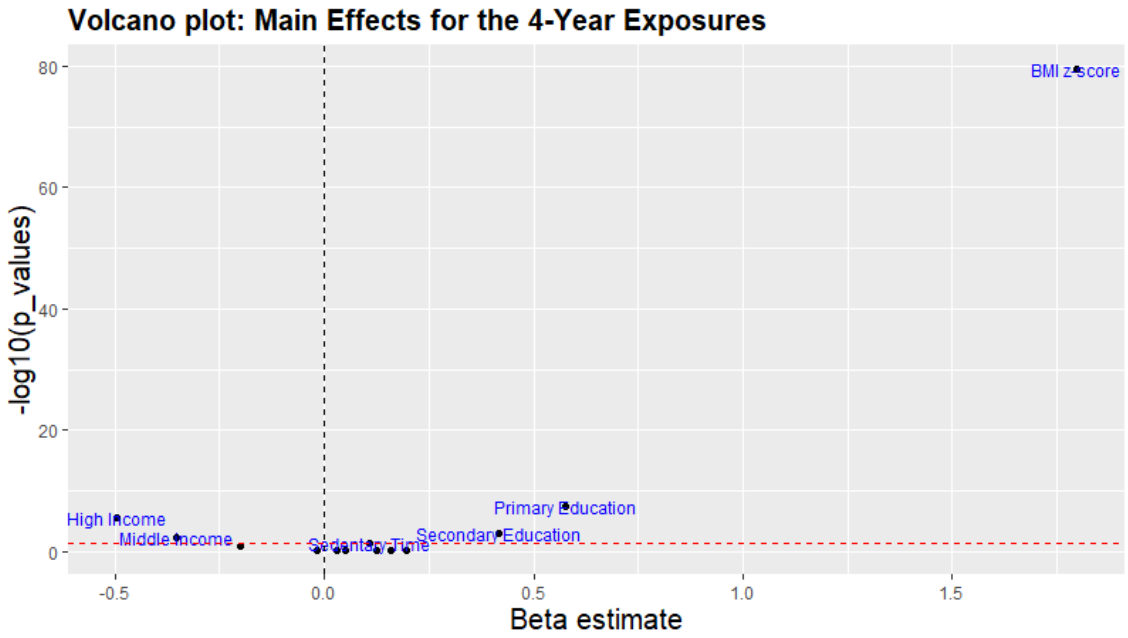


FIGURE 7.4: Associations between 4 y.o. exposures and the BMI z-score at age 13. The volcano plot shows the p-values against the beta coefficient. Red dashed horizontal line at the value of the FWER=0.05.

values. We see that the results for the logistic regression model, LDA, and Classification Tree are very similar. However, those from the logistic regression model have a straightforward interpretation (through odds ratio), which is advantageous, especially in a medical context. Therefore, the logistic regression model was selected.

Before fitting the model, the linear relationship between the continuous predictors and the outcome was assessed. Figure 7.5 does not seem to violate the linear assumption.

Final results for the logistic regression model are shown in table 7.3. We conclude that:

1. The odds of a child, whose mother has a primary education, being obese at age 13 are 57% greater than those having mothers with tertiary education.

2. The odds of a child, whose mother has a middle education, being obese at age 13 are 35% greater than those having mothers with tertiary education.

3. The odds of a child being obese at age 13 are 26% lower for those with a middle household income than children from low-income households.

4. The odds of a child being obese at age 13 are 29% lower for those with a high household income than children from low-income households.

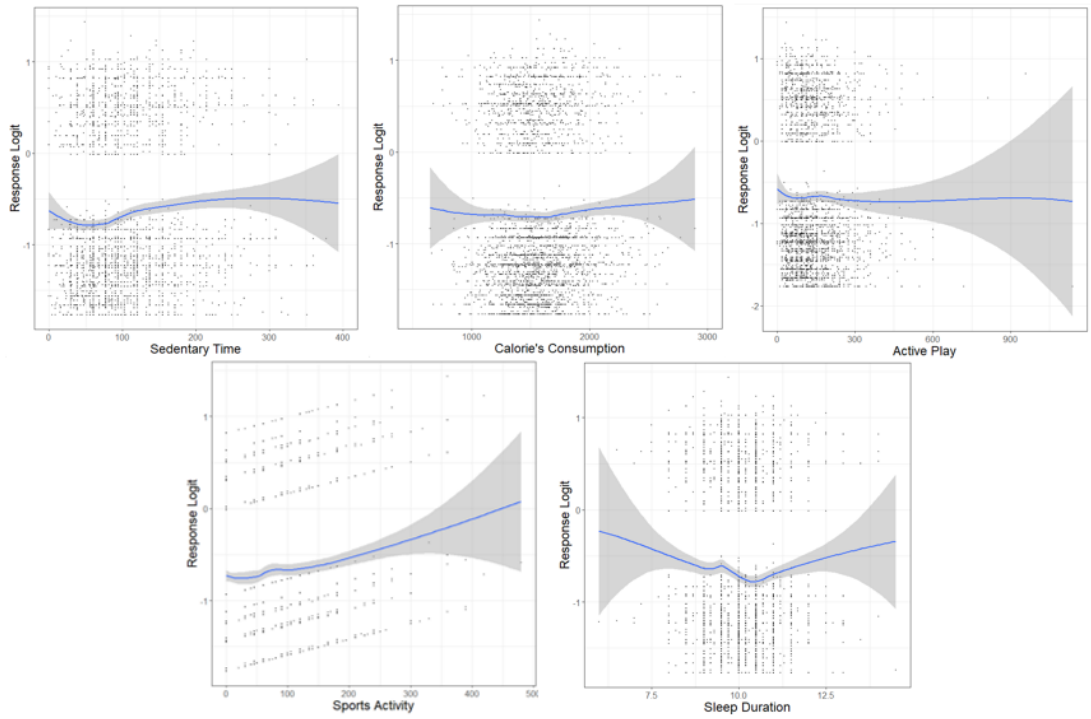


FIGURE 7.5: Scatter plot of the logit of the response and each continuous variable, for the 4 y.o. dataset. The regression line (in blue) was obtained by loess, and its confidence interval is pictured in grey.

TABLE 7.3: Odds Ratios and p-values for every variable present in the final 4 y.o.-model.

Exposures	**Odds-Ratio** (95% **CI)**	**p-value**
Education		
Primary	1.57 (1.21, 2.05)	< 0.001
Secondary	1.35 (1.05, 1.73)	0.016
Tertiary	Ref	Ref
Household Income		
Low	Ref	Ref
Middle	0.74 (0.58, 0.93)	0.011
High	0.71 (0.55, 0.91)	0.008
Z-score BMI		
Normal	Ref	Ref
Overweight/obese	6.00 (5.04, 7.16)	< 0.001
Hours of sports activity	1.10 (1.01, 1.19)	0.02

5. The odds of a child being obese at age 13 are 500% greater than those with a normal BMI at age 4.

6. For every hour of sports activity, the odds of a child being obese at age 13 increase by 10%. However, this value differs from what was expected. This difference could potentially be attributed to the fact that children who are already obese at the age of 4 are advised to engage in sports activities so a higher value for the hours practicing sports activity may be associated with overweight children.

7.1.3 Model for 7 years-old

The 7 y.o.-model includes 12 exposures that were collected at 7 y.o. of the child. Once again, before fitting the model with all exposures, the crude effects of each exposure were investigated. Table A.9 in Appendix A presents the estimated odds ratios and adjusted p-values. The exposures maternal education, household income, working condition, BMI, and child's sedentary time, soft drinks consumption, and child's BMI are significantly associated with the child being overweight/obese at age 13. Figure 7.8 shows a visual representation of the adjusted p-values. We can see that the BMI of the child had the most significant effect on the outcome.

Similar to the 4 y.o.-model, Logistic Regression, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Classification Trees were tested to choose the most suitable method. Table A.11 in Appendix A shows the results for those models' performance. Logistic Regression was selected.

Before fitting the model, the linear relationship between the continuous predictors and the outcome was assessed. Figure 7.7 does not seem to violate the linear assumption.

Final results for the logistic regression model are shown in table 7.4. We conclude that:

1. The odds of a child, whose mother has a primary education, being obese at age 13 are 34% greater than those having mothers with tertiary education.

2. Children from middle-income households have 27% lower odds of being obese at age 13 than those from low-income households.

3. Children from high-income households have 35% lower odds of being obese at age 13 than those from low-income households.

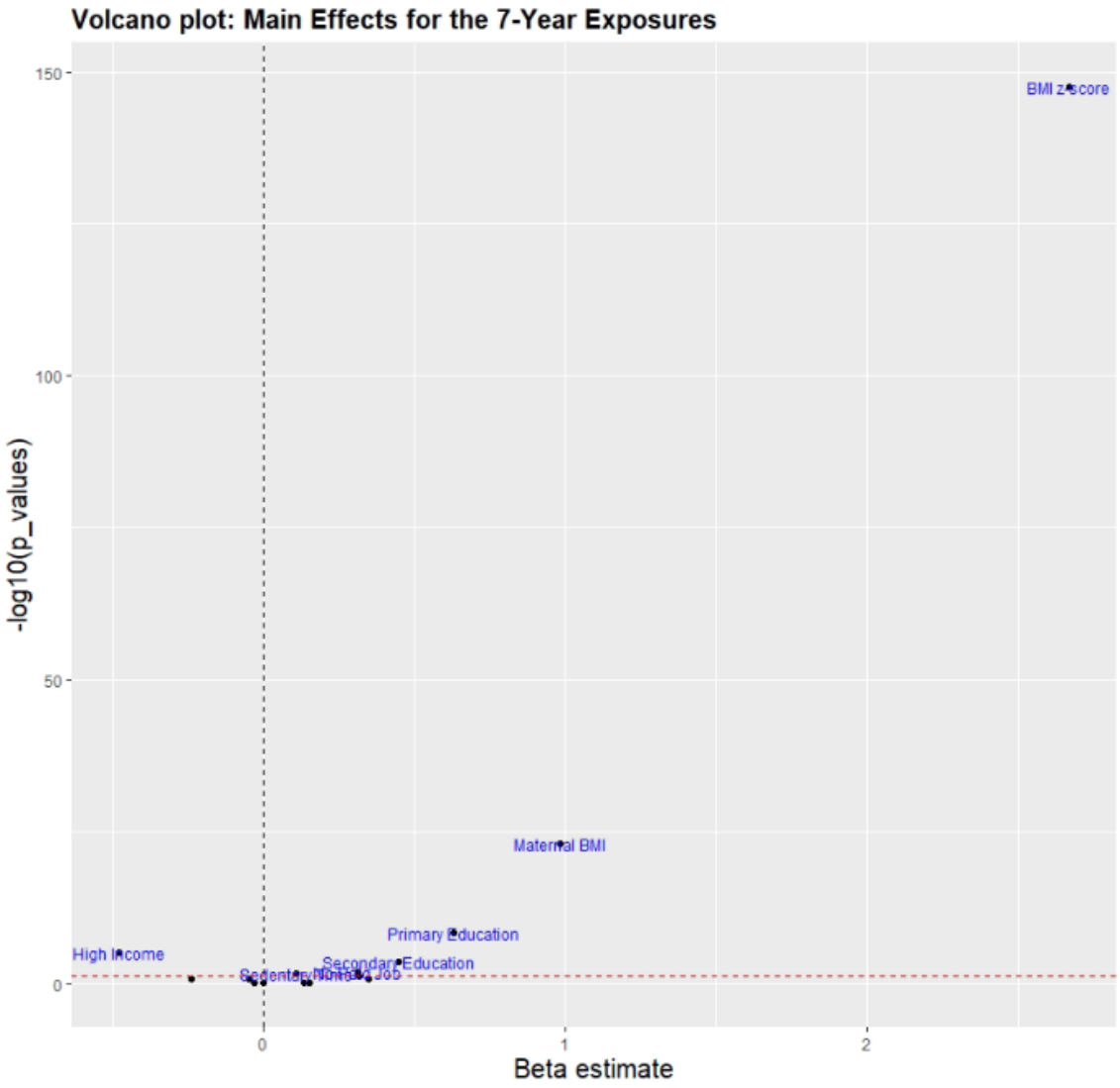


FIGURE 7.6: Associations between 7 y.o. exposures and the BMI z-score at age 13. The volcano plot shows the adjusted p-values against the beta coefficient. Red dashed horizontal line at the value of the FWER=0.05.

4. Children who are overweight or obese at 7 years old have significantly higher odds of being obese at age 13 than those with a normal Z-score BMI.

5. Children whose mothers are overweight or obese have significantly higher odds of being obese at age 13 than those whose mothers have a normal BMI.

6. Children who consume five or more portions of fruits, vegetables, or soup per day have 55% higher odds of being obese at age 13 than those who consume less than five portions per day. This value cannot be interpreted directly because, at age 7, only 7% of the observations consumed more than 5 portions a day so this variable is not correctly representing the sample.

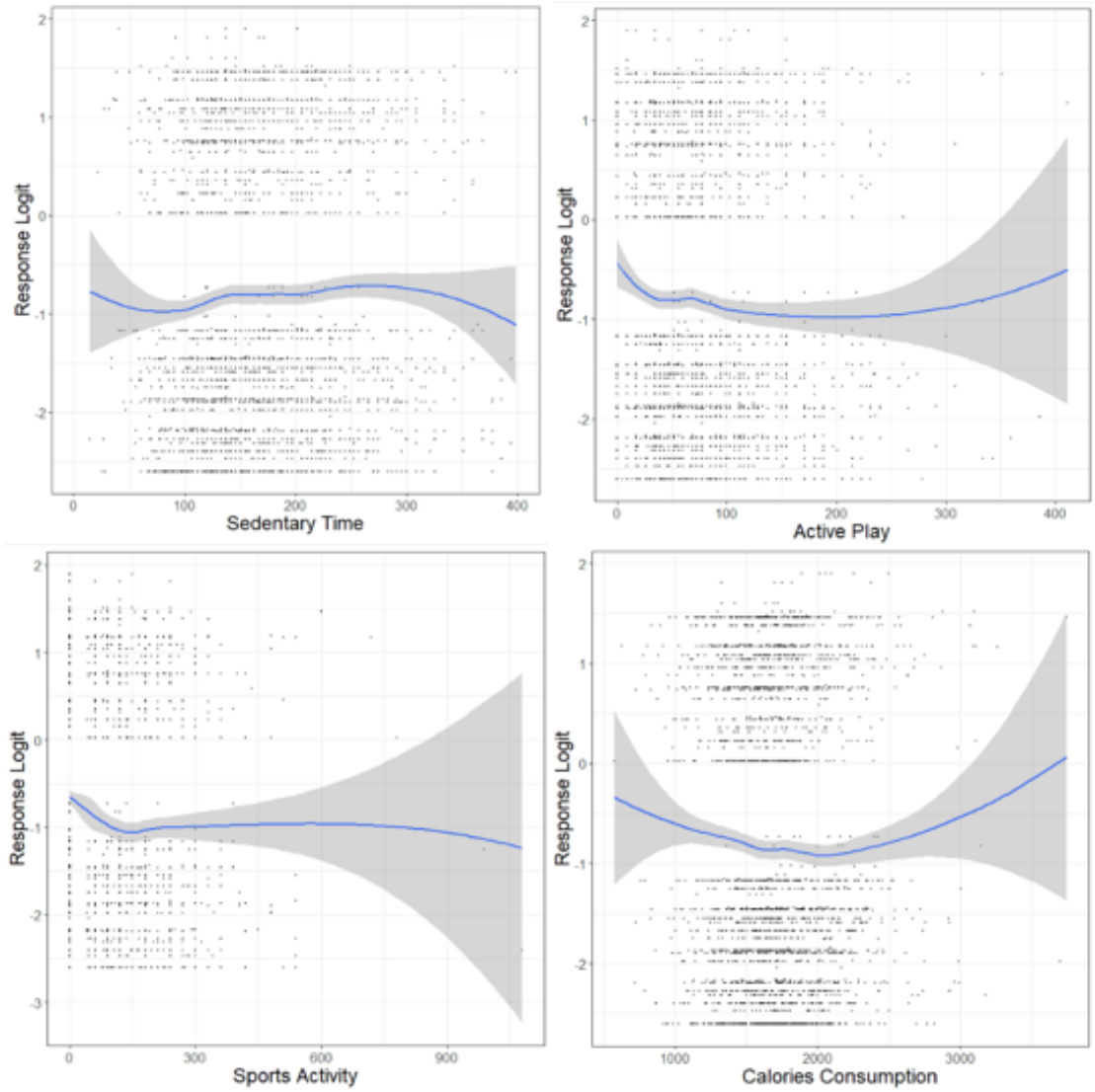


FIGURE 7.7: Scatter plot of the logit of the response and each continuous variable, for the 7 y.o. dataset. The regression line (in blue) was obtained by loess, and its confidence interval is pictured in grey.

7.1.4 Model for 10 years-old

The 10 y.o.-model includes 9 exposures that were collected at 10 y.o. of the child. Once again, before fitting the model with all exposures, the crude effects of each exposure were investigated. Table A.10 in Appendix A presents the estimated odds ratios and adjusted p-values. The exposures maternal education, household income, and child's sedentary time, soft drinks consumption, and child's BMI are significantly associated with the child being overweight/obese at age 13. Figure 7.8 shows a visual representation of the adjusted p-values. We can see that the BMI of the child had the most significant effect on the outcome.

TABLE 7.4: Odds Ratios and p-values for every variable present in the final 7 y.o.-model.

Exposures	**Odds-Ratio (95% CI)**	**p-Value**
Education		
Primary	1.34 (1.01, 1.78)	0.043
Secondary	1.23 (0.93, 1.60)	0.143
Tertiary	Ref	Ref
Household Income		
Low	Ref	Ref
Middle	0.73 (0.56, 0.95)	0.019
High	0.65 (0.50, 0.86)	0.002
Z-score BMI		
Normal	Ref	Ref
Overweight/obese	13.97 (11.50, 16.98)	< 0.001
Maternal BMI		
Normal	Ref	Ref
Overweight/obese	2.00 (1.61, 2.49)	< 0.001
Fruit/Vegetables/Soup		
$<$ 5 a day	Ref	Ref
$\geq$ 5 a day	1.55 (1.08, 2.26)	0.019

Once again, the methods of Logistic Regression, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Decision Trees were tested to choose the most suitable. Table A.11 in Appendix A shows the results for those models' performance. Logistic Regression was selected.

Before fitting the model, the linear relationship between the continuous predictors and the outcome was assessed. Figure 7.9 does not seem to violate the linear assumption

The final results for the logistic regression model are shown in table 7.5. The exposure maternal education was marginally significant but it was included in the model.

We conclude that:

1. The odds of a child, whose mother has a primary education, being obese at age 13 are 33% greater than those having mothers with tertiary education.

2. Children from high-income households have 33% lower odds of being obese at age 13 than those from low-income households.

3. Children who are overweight or obese at age 7 have significantly higher odds of being obese at age 13 than those with a normal Z-score BMI.

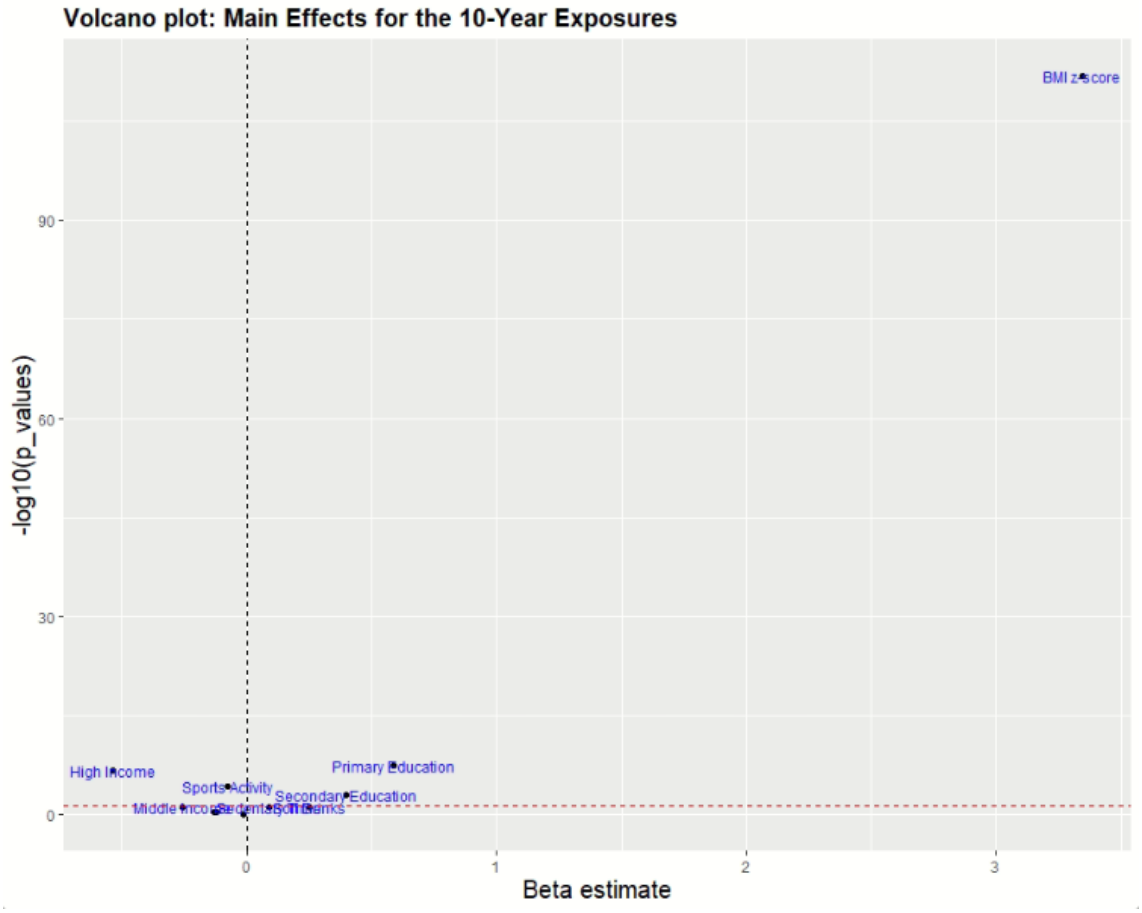


FIGURE 7.8: Associations between 10 y.o. exposures and the BMI z-score at age 13. The volcano plot shows the adjusted p-values against the beta coefficient. Red dashed horizontal line at the value of the FWER=0.05.

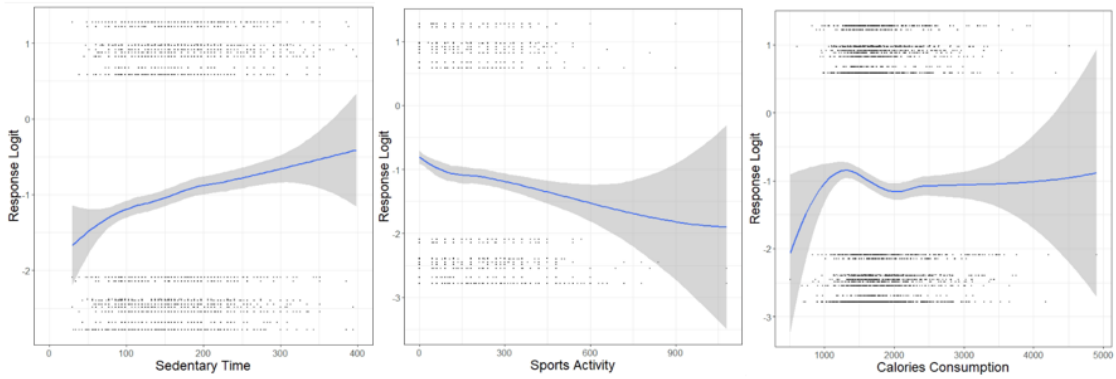


FIGURE 7.9: Scatter plot of the logit of the response and each continuous variable, for the 10 y.o. dataset. The regression line (in blue) was obtained by loess, and its confidence interval is pictured in grey.

TABLE 7.5: Results for the final logistic regression 10-year model with odds-ratio and corresponding confident intervals and P-values.

Exposures	**Odds-Ratio (95% CI)**	**p-Value**
Education		
Primary	1.33 (0.98, 1.81)	0.063
Secondary	1.22 (0.92, 1.61)	0.157
Tertiary	Ref	Ref
Household Income		
Low	Ref	Ref
Middle	0.77 (0.58, 1.01)	0.068
High	0.67 (0.50, 0.91)	0.010
Z-score BMI		
Normal	Ref	Ref
Overweight/obese	28.12 (22.38, 35.33)	< 0.001

7.1.5 Models' Performance

The performance of the previous models was evaluated by the following measures:

- Sensitivity: measures the proportion of actual positive cases (children obese at age 13) correctly identified by the model. It is calculated as the number of true positives divided by the sum of true positives and false negatives. Sensitivity quantifies how well a model detects positive cases.

- Specificity: measures the proportion of actual negative cases (children with normal weight at age 13) correctly identified by a model. It is calculated as the number of true negatives divided by the sum of true negatives and false positives. Specificity quantifies how well a model identifies negative cases.

- AUC: represents the performance of a classification model across all possible classification thresholds. It measures the ability of the model to distinguish between positive and negative classes. AUC ranges from 0 to 1, a higher value indicates better performance.

- Positive Predictive Ability (PPA): measures the proportion of predicted positive cases that are truly positive. It is calculated as the number of true positives divided by the sum of true positives and false positives. PPA quantifies how well a model correctly predicts positive cases.

- Negative Predictive Ability (NPA): measures the proportion of predicted negative cases that are truly negative. It is calculated as the number of true negatives divided by the sum of true negatives and false negatives. NPA quantifies how well a model correctly predicts negative cases.

- Prediction Error: measures the proportion of incorrect predictions made by the classification model.

All measures, except for the AUC, are computed from the confusion matrix, corresponding to the cross-table between the model's predictions and the (binary) response variable. It is important to notice that the model returns the probability of a certain child being obese at age 13 but, in the confusion matrix a cut-off value was defined and the observations with probabilities predicted by the model below that cut-off were assigned to the non-obese class and the observations with probabilities above the cut-off were assigned to the obese class. The cut-off value was calculated by the maximum Younden index (sensitivity+specificity-1).

TABLE 7.6: Performance measures for each model.

Models	Specificity	Sensitivity	AUC	PPA	NPA	Prediction Error
Pregnancy model	0.71	0.61	0.70	0.53	0.77	33 %
Pregnancy Model (ENET)	0.75	0.57	0.71	0.49	0.80	36 %
4-Year Model	0.86	0.61	0.76	0.71	0.80	23 %
7-Year Model	0.86	0.67	0.82	0.70	0.84	20%
10-Year Model	0.88	0.81	0.87	0.70	0.93	17 %
Dynamic Model ENET	0.86	0.81	0.88	0.72	0.91	17 %

The results are shown in table 7.6. These values were obtained by fitting the final model in each imputed dataset. Then the performance measures were calculated for every dataset and the mean of each measure was considered. Firstly, the table shows that the performance measures for the two models applied to the pregnancy dataset provide very similar values, although the prediction error of the first model is slightly lower. Secondly, we can see that the largest prediction error is for the model that only includes variables about the pregnancy stage and the first two years of life, which is expected. In fact, table 7.6 demonstrates a decrease in prediction error, accompanied by an increase in AUC and specificity/sensitivity when exposures collected closer to the outcome age (13) are included in the model. Lastly, we conclude that all models predict better children who will not be obese than those that will.

By looking at the 3 models related to the child (4 y.o.-model, 7 y.o.-model, and 10 y.o.-model) we see that the child's BMI z-score is the predictor that mostly influences the risk of becoming obese at 13. Moreover, the odds ratio of this variable increases as the child's age considered in the model increases. We thus raised two questions: (1) do models that in each time point include only the child's BMI z-score perform worse than the complete model presented before?; (2) what happens if the child's BMI is excluded from the regression?

We fitted new models and concluded that:

1. The performance measures of each new model (with only the child's BMI z-score) were nearly identical to the previous models, with only a slight decrease in AUC values. This strongly suggests that prevention should rely on the child's BMI during childhood.

2. The models had a serious decrease in each performance measure. The prediction errors were of 43 %, 42 %, and 38 % for the 4 y.o., 7 y.o., and 10 y.o. models, respectively. Additionally, new associations were identified by the final model. For the 4 y.o.-model, the sedentary time was significant with an odds-ratio value of 1.09 (for a 1-hour increase in the hours of sedentary time, the odds of being obese at age 13 was 9% higher). For the 10 y.o.-model, the hours of sports activity became significant with an odds ratio of 0.95 (for a 1-hour increase in the hours of sports activity the odds of being obese at age 13 decreased by 5%).

7.2 Dynamic Model

The dynamic model was fitted considering all exposures at once, using 2 models: (1) a penalized regression model; (2) a finite mixture of penalized regressions.

7.2.1 Penalized Regression Model

In section 3.3, three different penalizations were presented: LASSO, Ridge, and ENET. Since the dynamic model included all exposures, we thought ENET would be the most suitable penalization since it can handle a large number of correlated predictors while performing both feature selection and regularization. However, a straightforward application of the traditional ENET regression could not be performed since the MICE algorithm returned 10 imputed datasets (m=10). Therefore, the method described in section

3.4 was applied by considering a stacked objective function. The model was fitted by using the function *cv.saenet* from the R-package *miselect*. This function performs 5-fold cross-validation to select the optimal values for λ and α for the elastic net regression model.

The model was fitted on the 55 available exposures. The optimal value for λ was found to be 0.009 and for α was 0.1. The value of λ controls the overall level of regularization applied by the elastic net model. Therefore, the optimal value of 0.009 indicates that a low level of regularization was applied to the model. The value of α determines the balance between the L1 (LASSO) and L2 (Ridge) regularization terms in the elastic net model. A value of 0.1 suggests that the model primarily relies on L1 regularization, indicating a preference for sparse solutions with a smaller number of important features. The estimated odds ratio are shown in table 7.7. Once again, the confidence intervals for the odds ratio were not computed. Among the 55 variables, the algorithm selected 27.

Looking at the table, we see that there are odds ratios equal to 1, meaning that the corresponding exposure does not have an effect on the response. Moreover, we note a significant reduction in the odds ratios compared to those obtained for the static models. This substantial decrease is attributed to the inclusion of a penalization term in the analysis. Additionally, we once again identify the child's z-score BMI as the predictor with the largest effect on obesity at age 13, as its effect increases as the time-point increases (the odds ratio at age 4 was estimated at 1.082 but was 1.464 at age 10).

The corresponding performance measures are presented in table 7.6. The values are very similar to the ones obtained for the 10 y.o. static model.

7.2.2 Finite Mixtures of Regressions

We started by fitting the model with a binary outcome, corresponding to the BMI status at 13 years old (overweight/obese versus normal). As the algorithm did not converge, we considered instead the continuous BMI. The function *stepflexmix*, from the R-package *flexmix*, was used to run the model and select the optimal number of components. As the package does not allow the fitting on multiply imputed datasets, we consider the first complete dataset out of the imputed m. The finite mixture of regressions was performed using penalized regressions with adaptive LASSO regularization. The analysis was conducted with a number of components ranging from 1 to 15. The optimal number of clusters was based on the Bayesian Information Criterion (BIC). Figure 7.10 displays the values of the criteria according to the number of clusters. We see that a 2-component

TABLE 7.7: Odds-ratio for the final elastic net model.

Exposures	**Odds-Ratio**			
	Pregnancy	4 y.o.	7 y.o.	10 y.o.
Age	0.998	-	-	-
Education				
Primary	1.005	-	1.000	1.008
Secondary	1.000	-	1.004	1.000
Tertiary	Ref	Ref	Ref	Ref
Working Condition				
Paid job	Ref	-	Ref	-
No paid job activity	1.011	-	1.020	-
Marital Status				
Married or cohabiting	Ref	Ref	Ref	Ref
No partner	1.028	1.009	1.017	-
Household Income				
Low	Ref	Ref	Ref	Ref
Middle	1.000	0.986	0.991	-
High	0.965	0.992	1.000	-
Parity				
Nullipara	Ref	-	-	-
Multipara	1.019	-	-	-
Smoking Habits during Pregnancy				
Never smoked	Ref	-	-	-
Smoked	1.016	-	-	-
Gestational Diabetes				
Yes	1.013	-	-	-
No	Ref	-	-	-
Maternal BMI				
Underweight/normal	Ref	-	Ref	-
Overweight/obese	1.040	-	1.170	-
Breastfeeding				
Never	1.007	-	-	-
0-2.9 months	0.994	-	-	
3-5.9 months	0.985	-	-	
$\geq$ 6 months	Ref	-	-	
Age of first solid food				
$<$ 4 months	1.032	-	-	
4-5 months	Ref	-	-	
$\geq$ 6 months	1.019	-	-	
BMI 2 years	1.007	-	-	-
Z-score BMI				
Normal	-	Ref	Ref	Ref
Overweight/obese	-	1.082	1.170	1.464
Sedentary time	-	-	-	1.000
Hours of sports activity	-	1.000	-	1.000
Calories consumption	-	1.000	-	-
Fruit/Vegetables/Soup				
$<$ 5 a day	-	Ref	Ref	Ref
$\geq$ 5 a day	-	-	1.010	-

structure provides the most suitable representation of the data since it corresponds to the lowest value.

Cluster 1 has 2187 observations, and Cluster 2 has 796. The first component of the model has a prior probability of 0.6 and the second component has a prior probability of 0.4. In a finite mixture model, each data point is assigned a probability of belonging to each cluster. These probabilities are called posterior probabilities and are estimated based on the observed data and the model parameters. Figure 7.11 presents the histograms for

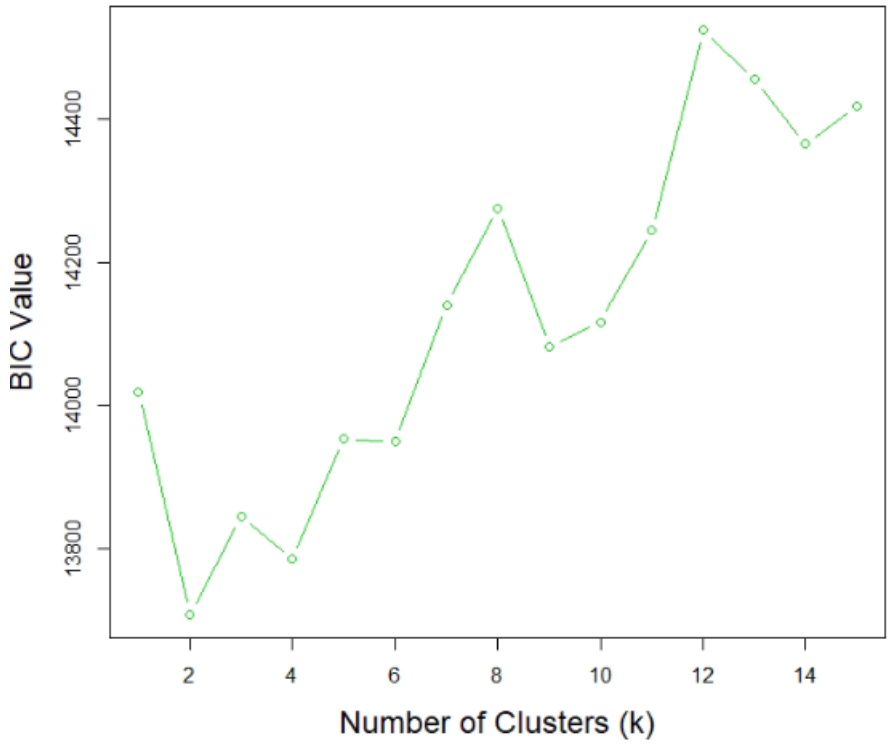


FIGURE 7.10: BIC Values for each number of clusters.

the posterior probabilities for each cluster. The plot shows that there are several observations with a posterior probability near 0.5 which means that there is a certain overlap between the two components. The mean (standard deviation) of the posterior probabilities from Cluster 1 was 0.75 (0.12) and 0.83 (0.17) for Cluster 2, which are satisfactory values.

The choice of 2 clusters indicates the presence of two distinct subgroups within the data, characterized by different regression models. Table 7.9 shows the estimated coefficients for the regression model in each cluster. The cells that are not filled in correspond to the coefficients that were shrunk to 0.

For a better understanding of the clusters, the dataset was divided into two datasets, and a descriptive analysis of the variables was performed. Dataset i with $i = 1, 2$ included the observations assigned to cluster i. The following was concluded:

1. **Cluster 1 contains**: mothers with higher school education (33% with higher education), mothers with higher household income (38% with higher education), more mothers that did not report gestational diabetes or gestational hypertensive complications in comparison with the observations from cluster 2, mothers with slightly

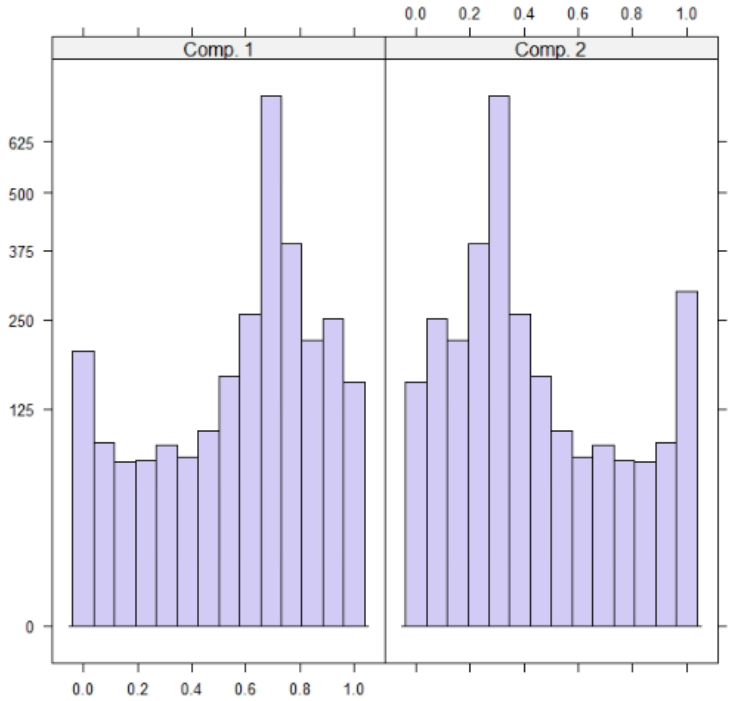


FIGURE 7.11: Histogram of the posterior probabilities in each cluster.

lower BMI value (mean equal to 23.5), more mothers with normal weight gain during pregnancy (40 %), more boys than girls (52 %), more children with a normal BMI z-score (74 % at 4 y.o., 68 % at 7 y.o., 65 % at 10 y.o.), and contains children with less BMI at 13 y.o. (mean=19.32).

2. **Cluster 2 contains**: mothers with lower school education (26% with higher education), mothers with lower household income (30% with higher education), mothers with slightly lower BMI value (mean equal to 25.0), more mothers with excessive weight gain during pregnancy (42 %), more girls than boys (53 %), fewer children with a normal BMI z-score (54 % at 4 y.o., 45 % at 7 y.o., 38 % at 10 y.o.), and contains children with higher BMI at 13 y.o. (mean=25.00).

The evaluation of the model's performance made use of the following measures:
Mean Squared Error (MSE):

$$MSE = \sum_{i=1}^{n} (\text{observed value} - \text{predictive value})^2 \tag{7.1}$$

Mean Absolute Error (MAE):

$$MAE = \sum_{i=1}^{n} |\text{observed value} - \text{predictive value}| \tag{7.2}$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \sum_{i=1}^{n} \left| \frac{\text{observed value} - \text{predictive value}}{\text{observed value}} \right| \times 100 \tag{7.3}$$

These measures were calculated in two ways:

1. Weighted Mean Prediction: for each observation, we used the model estimated in each cluster to predict the outcome. The final predictive value resulted from a weighted mean of the two predictions. The weights were determined by multiplying each prediction by the posterior probability of that observation belonging to each cluster.

2. Maximum Posterior Probability Prediction: the prediction was based solely on the model of the component providing the maximum posterior probability. We selected the prediction from this model as the final predictive value.

The results for each method are shown in table 7.8. The results are very satisfactory with a MAPE value of only 5.6%. This means that the model is suitable to predict the child's BMI at age 13.

Since this finite mixture of regressions was fitted considering the outcome as continuous, it is not possible to compare the performance of the ENET regression model with that of the finite mixture of regressions. To address this limitation, I re-fitted the ENET regression solely on the first complete imputed dataset, which was used for applying the finite mixture model. In this case, the outcome variable was treated as continuous. Table 7.8 shows the results. It is possible to conclude that the finite mixture model when the weighted mean method is used to predict the values, performs better than the ENET regression.

TABLE 7.8: Performance measures for the dynamic models.

Model	**MSE**	**MAE**	**MAPE**
Finite Mixture Model (Weighted Mean)	2.60	1.15	5.6 %
Finite Mixture Model (Maximum Posterior Probability)	8.21	2.03	8.9 %
ENET	5.87	1.85	8.63%

TABLE 7.9: Coefficients for the model in each cluster.

Exposures	**Cluster 1**				**Cluster 2**			
	Pregnancy	4 y.o.	7 y.o.	10 y.o.	Pregnancy	4 y.o.	7 y.o.	10 y.o.
Education								
Primary	0.266	-0.582	0.490	-0.376	-	0.000	-	0.000
Secondary	0.169	0.000	0.000	0.000	-	-0.141	-	0.265
Tertiary	Ref	Ref	Ref	Ref	-	Ref	-	Ref
Marital Status								
Married or cohabiting	Ref	-	Ref	-	Ref	Ref	-	-
No partner	0.205	-	0.057	-	0.504	-0.442	-	-
Household Income								
Low	Ref	Ref	Ref	Ref	Ref	Ref	-	Ref
Middle	0.000	-0.024	-0.194	-	-	-0.346	-	0.000
High	0.0632	0.000	-0.131	-	-	-0.775	-	-0.064
Smoking Habits during Pregnancy								
Never smoked	Ref	-	-	-	Ref	-	-	-
Smoked	0.185	-	-	-	0.274	-	-	-
Gestational Diabetes								
Yes	-0.009	-	-	-	-	-	-	-
No	Ref	-	-	-	-	-	-	-
Maternal BMI	0.055	-	-	-	0.143	-	-	-
Ceaserean mode of delivery								
Yes	0.017	-	-	-	0.0005	-	-	-
No	Ref	-	-	-	Ref	-	-	-
Vaginal mode of delivery								
Yes	-	-	-	-	0.244	-	-	-
No	-	-	-	-	Ref	-	-	-
Breastfeeding								
Never	0.000	-	-	-	0.172	-	-	-
0-2.9 months	0.000	-	-	-	0.000	-	-	-
3-5.9 months	-0.087	-	-	-	0.000	-	-	-
$\geq$ 6 months	Ref	-	-	-	Ref	-	-	-
Age of first solid food								
$<$ 4 months	0.286	-	-	-	-	-	-	-
4-5 months	Ref	-	-	-	-	-	-	-
$\geq$ 6 months	-0.07	-	-	-	-	-	-	-
First solid food								
Cereal porridge	0.000	-	-	-	-0.327	-	-	-
Soup	Ref	-	-	-	Ref	-	-	-
Fruit	0.131	-	-	-	-0.41	-	-	-
Gestational weight gain								
Insufficient	-0.194	-	-	-	-	-	-	-
Normal	Ref	-	-	-	-	-	-	-
Excessive	0.000	-	-	-	-	-	-	-
BMI 6 months	0.067	-	-	-	-	-	-	-
BMI 2 years	0.085	-	-	-	0.104	-	-	-
Z-score BMI								
Normal	-	Ref	Ref	Ref	-	Ref	Ref	Ref
Overweight/obese	-	0.447	0.944	2.385	-	1.53	1.89	3.992
Newborn's Sex								
Boy	-0.854	-	-	-	-1.032	-	-	-
Girl	Ref	-	-	-	Ref	-	-	-
Newborn's Length	-0.027	-	-	-	-	-	-	-
Sleep Duration	-	-0.051	-	-	-	-	-	-
Fruit/Vegetables/Soup								
$<$ 5 a day	-	Ref	Ref	Ref	-	Ref	-	-
$\geq$ 5 a day	-	0.265	0.069	-0.100	-	-0.216	-	-
Soft drinks								
never or less than 1 per week	-	-	-	-	-	-	-	Ref
1-6 per week	-	-	-	-	-	-	-	0.149
one or more per day	-	-	-	-	-	-	-	0.000

Chapter 8

Conclusions

The objective of this internship was to develop a dynamic prediction model for childhood obesity considering an exposome approach. Childhood obesity is a global health crisis, with significant implications for the physical and mental well-being of children and potential long-term impacts on public health systems. So the goal was to build prediction models that could be applied in clinical contexts to ensure early prevention.

Our findings indicate a strong association between several exposures collected during pregnancy and childhood, and obesity at 13 y.o., including household income, maternal education, maternal body mass index before pregnancy, smoking habits during pregnancy, a child's sedentary time, and child BMI, among others. In fact, the results show that the Child's BMI measured at each follow-up was systematically the most important predictor for obesity at 13 y.o.. The four static models presented earlier enable us to understand that the variables regarding the pregnancy stage of the mother, only, are not enough to accurately predict obesity. On the opposite, considering only exposures collected at 10 y.o. provided 17% of prediction error, which is a satisfactory value. So, due to its simplicity, the 10 y.o.-model can be more advantageous to predict childhood obesity than the dynamic model with ENET regression, which had the same prediction error. The dynamic prediction model considering finite mixture models demonstrated high accuracy in forecasting obesity risk, highlighting its potential utility in early intervention strategies. However, this method can be more difficult to implement in a clinical context.

In conclusion, this report offers a significant step forward in predicting childhood obesity risk using a dynamic, early-life exposome approach. The model developed in this research can potentially change the landscape of early detection and prevention of childhood obesity, offering a more accurate and comprehensive tool for risk prediction.

Besides the epidemiologic perspective, this thesis also conducted a relevant analysis on addressing missing values. The two presented simulation studies led to the following conclusions:

1. Among the six tested imputation methods, Multiple Imputations with Chained Equations using predictive mean matching, logistic regression, and polytomous regression as the imputation models for continuous, binary, and categorical variables, respectively, exhibited the closest estimated logistic regression coefficient values to those obtained from the original dataset.

2. The relative efficiency of the imputation procedure remained consistent across the three imputation models considered for Multiple Imputation by Chained Equations.

3. MICE with linear models, logistic regression, and polytomous regression as the imputation models for continuous, binary, and categorical variables, respectively, resulted in the imputed datasets with the highest concordance measures.

8.1 Future Work

This report focused approached missing data problems, however, measurement error, is also a common challenge in data analysis, particularly in self-reported questionnaires which is the case of the Generation XXI cohort. There are several methods to deal with measurement error. One particularly interesting method is to treat measurement error as a missing data problem. So future work could be to adapt some of the presented methodologies to a measurement error context. In this framework, measurement errors are seen as partially missing information and completely missing values as an extreme form of measurement error.

Although three domains for the exposome were presented, the work focuses only on the general external exposome and also included some exposures related to the specific external exposome. In future work, the prediction of childhood obesity can also be performed including the urban exposome and the internal domain exposome.

Finally, the R-package *flexmix*, which runs finite mixture models, does not allow the function to be applied to a context of multiply imputed datasets. In fact, in this thesis, only one of the imputed datasets was considered. Additionally, although the package allows for penalized regression methods, the only penalization available is the adaptive LASSO.

Thus, future work should also include the extension of this package to ENET penalization and to the fitting on multiple imputed datasets.

Appendix A

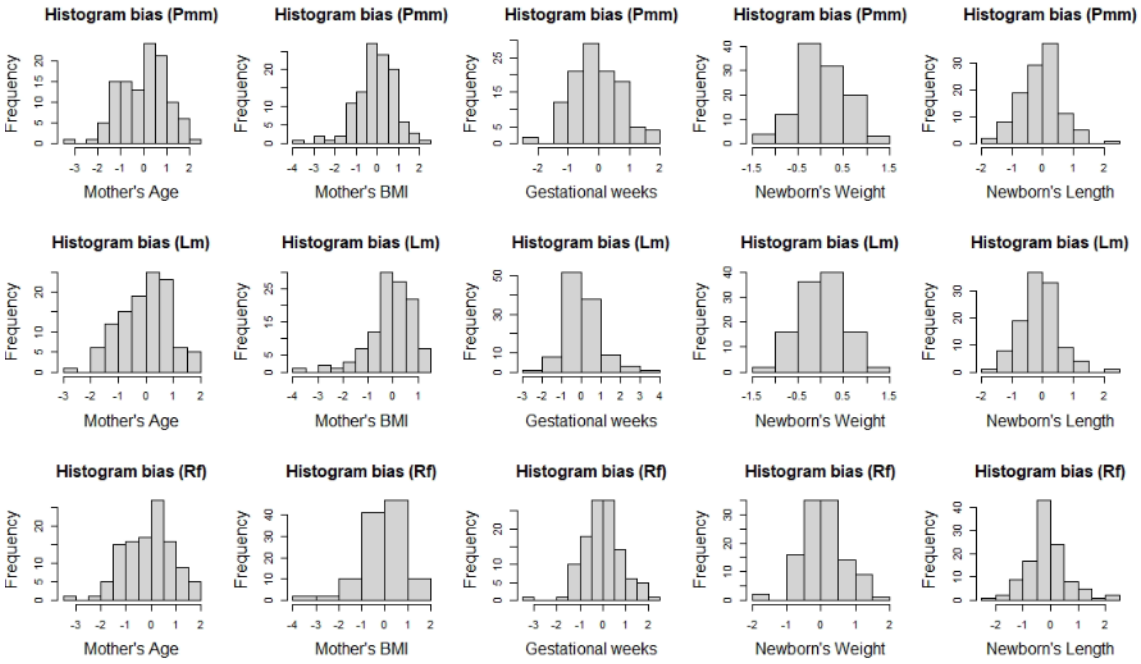


FIGURE A.1: Histograms for the bias for each method and variable.

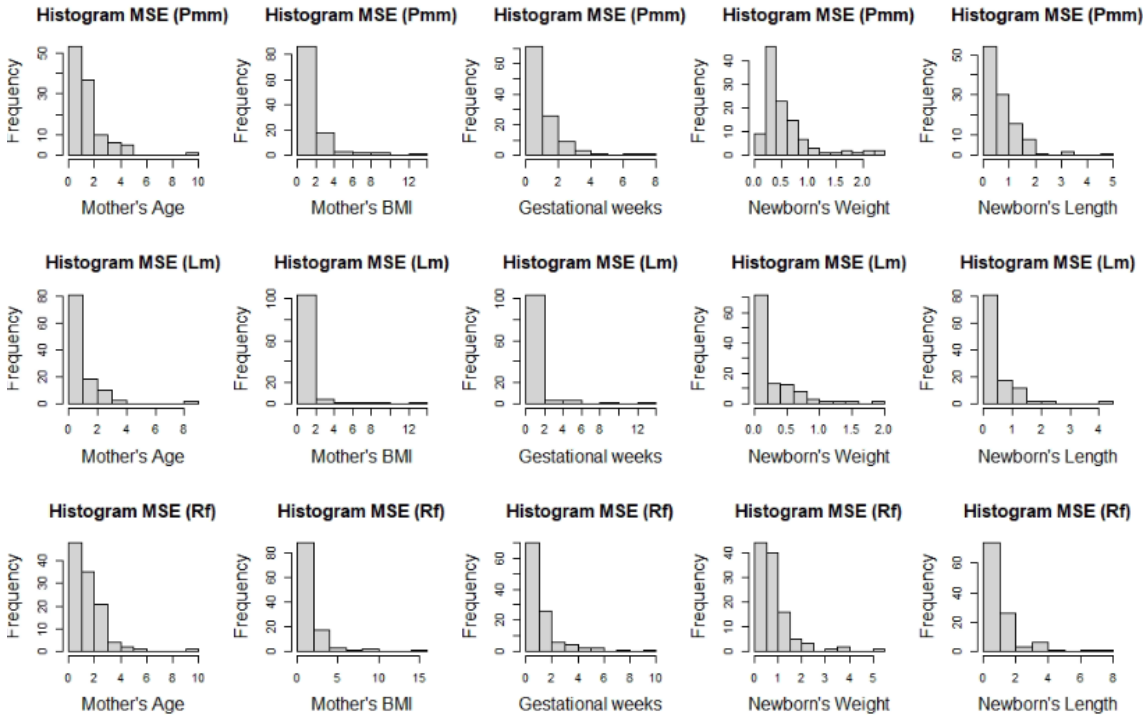


FIGURE A.2: Histograms for the Mean Squared Error (MSE) for each method and variable.

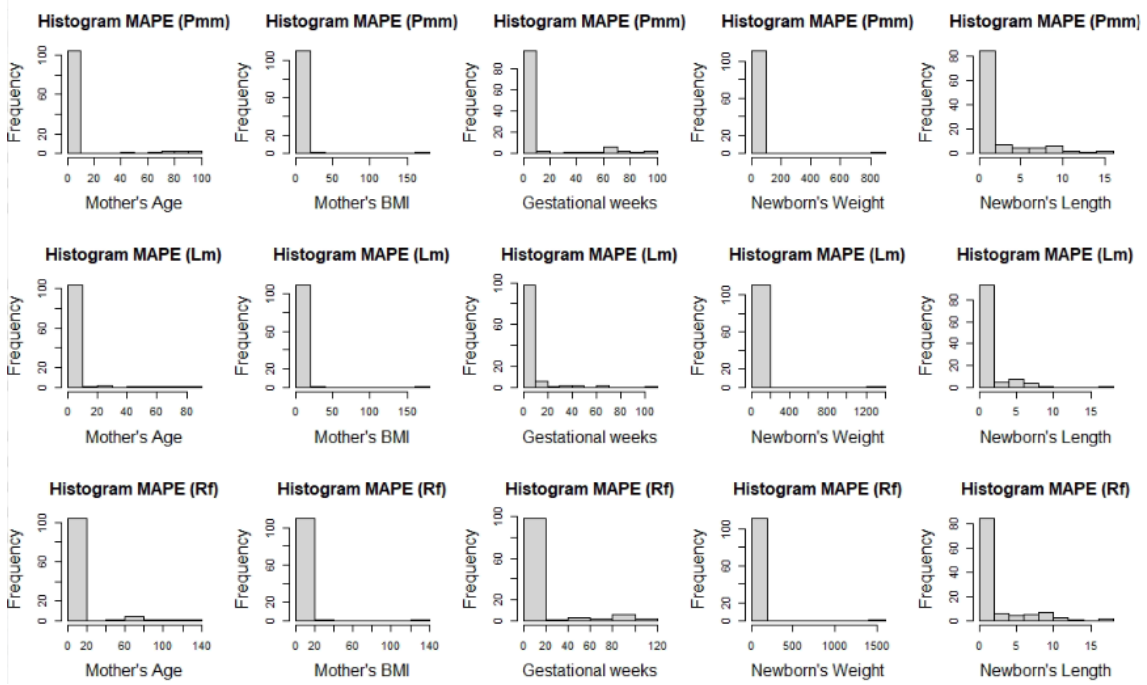


FIGURE A.3: Histograms for the Mean Absolute Percentage Error (MAPE) for each method and variable.

TABLE A.1: Descriptive analysis for the pregnancy/infancy variables.

Exposures	Obese	Non-obese
Sociodemographic and Psychosocial factors		
Age mean (sd) (years)	30 (5.32)	30 (4.98)
Education, n (%)		
Primary	691 (49 %)	995 (38%)
Secondary	398 (28 %)	708 (27 %)
Tertiary	331 (23 %)	887 (34 %)
Working Condition, n (%)		
Paid job	1124 (76 %)	2189 (81 %)
No paid job activity	361 (24 %)	516 (19 %)
Marital Status, n (%)		
Married or cohabiting	1415 (95%)	2631 (96 %)
No partner	82 (5%)	107 (4 %)
Household Income, n (%)		
Low	495 (38%)	738 (30 %)
Middle	422 (32%)	742 (31 %)
High	396 (30 %)	943 (39 %)
Parity, n (%)		
Nullipara	844 (57%)	1594 (59 %)
Multipara	638 (43%)	1105 (41 %)
Baby's Sex, n (%)		
Boy	730 (49 %)	1335 (49 %)
Girl	769 (51 %)	1412 (51 %)
Anthropometric factors		
Pre-conception BMI, n (%)		
Underweight/normal	781 (56%)	1935 (76 %)
Overweight/obese	613 (44%)	613 (24 %)
Gestational weight gain, n (%)		
Insufficient	275 (21%)	619 (26 %)
Normal	452 (34%)	981 (41 %)
Excessive	605 (45 %)	812 (34 %)
Baby's Weight mean (sd) (g)	3250 (484.2)	3200 (483.44)
Baby's Length mean (sd) (cm)	49 (2.16)	49 (2.38)
BMI 6 months mean (sd)	17.58 (1.51)	16.99 (1.43)
BMI 1 year mean (sd)	17.85 (1.50)	17.31 (1.41)
BMI 2 years mean (sd)	17.06 (1.57)	16.28 (1.25)
Lifestyle behaviors		
Smoking Habits during Pregnancy, n (%)		
Never smoked	1148 (78%)	2243 (83 %)
Smoked	327 (22 %)	465 (17%)
Breastfeeding, n (%)		
Never	411 (30 %)	623 (25 %)
0-2.9 months	260 (19%)	497 (20%)
3-5.9 months	226 (16%)	495 (20 %)
$\geq$ 6 months	489 (35%)	923 (36%)
Age of first solid food, n (%)		
<4 months	112 (9%)	150 (6%)
4-5 months	744 (60 %)	1507 (64%)
$\geq$ 6 months	393 (31 %)	710 (30%)
First solid food, n (%)		
Cereal porridge	693 (54%)	1470 (61%)
Fruit	80 (6%)	133 (5%)
Soup	522 (40 %)	823 (34 %)
Clinical Exposures		
Hypertensive complications, n (%)		
Yes	179 (12%)	229 (8 %)
No	1316 (88%)	2498 (92 %)
Gestational Diabetes, n (%)		
Yes	134 (9 %)	159 (6 %)
No	1361 (91 %)	2568 (94 %)
Vaginal Delivery, n (%)		
Yes	881 (60 %)	1700 (64 %)
No	581 (40 %)	971 (36 %)
Cesarean delivery, n (%)		
Yes	607 (42 %)	1005 (38 %)
No	844 (58%)	1623 (62 %)
Gestational weeks mean (sd)	39.5 (1.59)	39.51 (1.64)

TABLE A.2: Descriptive analysis for the pre-school/school variables. NA: Variable not available at that time point.

Exposures	**4 Years**		**7 Years**		**10 Years**	
	Obese	**Non-obese**	**Obese**	**Non-obese**	**Obese**	**Non-obese**
Sociodemographic						
Education, n (%)						
Primary	641 (46%)	914 (36 %)	571 (42%)	823 (33%)	290 (22%)	410 (17%)
Secondary	404 (29%)	712 (28%)	432 (32%)	740 (30%)	324 (24%)	483 (19%)
Tertiary	350 (25%)	910 (36%)	354 (16%)	926 (37%)	728 (54%)	1588 (64%)
Working Condition, n (%)					NA	NA
Paid job	1142 (78 %)	2170 (81 %)	1099 (77%)	2134 (81%)		
No paid job activity	325 (22 %)	504 (19 %)	327 (23%)	495 (19%)		
Marital Status, n (%)					NA	NA
Married or cohabiting	1293 (89%)	2422 (91%)	1229 (86%)	2333 (89%)		
No partner	153 (11%)	225 (9%)	199 (14%)	298 (11%)		
Household Income, n (%)						
Low	426 (30%)	558 (21%)	451 (32%)	611 (23%)	428 (31%)	581 (23%)
Middle	425 (29%)	760 (29%)	404 (28%)	745 (28%)	421 (30%)	733 (28%)
High	590 (41%)	1311 (50%)	572 (40%)	1270 (48%)	545 (39%)	1263 (49%)
Anthropometric factors						
Maternal BMI, n (%)	NA	NA			NA	NA
Underweight/normal			393 (30%)	1292 (53%)		
Overweight/obese			900 (70%)	1131 (47%)		
Z-score BMI, n (%)						
Normal	0 (0%)	2000 (83%)	324 (24%)	2076 (81%)	177 (13%)	2065 (81%)
Overweight/obese	1499 (100%)	414 (17%)	1052 (76%)	483 (19%)	1204 (87%)	499 (19%)
Lifestyle behaviors						
Sedentary time						
median (min-max) (min/day)	94 (0-497)	90 (0-677)	163 (27.14-540)	159 (3.57-505.71)	154 (30-471)	154 (30-454)
Active play					NA	NA
median (min-max) (min/day)	120 (0-497)	120 (0-677)	69 (0-411.43)	73 (0-385.71)		
Hours of sports activity						
median (min-max) (min/week)	60 (0-480)	60 (0-540)	0 (0-1080)	60 (0-990)	90 (0-1080)	120 (0-1260)
Sleep duration, mean (sd) (hours)	10 (0.92)	10 (0.80)	NA	NA	NA	NA
Fruit/Vegetables/Soup, n (%)						
$<$ 5 a day	571 (46%)	956 (41%)	1242 (92%)	2330 (93%)	1131 (83%)	2025 (80%)
$\geq$ 5 a day	680 (54%)	1372 (59%)	11 (8.2%)	183 (7%)	235 (17%)	514 (20%)
Calories consumption						
median (min-max) (kcal)	1560 (542-2885)	1554 (480-2946)	1678 (579-3742)	1696 (627-3696)	1713 (603-4900	1759 (513-4209)
Soft drinks, n (%)	NA	NA				
never or less than 1 per week			220 (16%)	509 (20%)	227 (17%)	496 (19%)
1-6 per week			687 (50%)	1255 (49%)	811 (59%)	1372 (54%)
one or more per day			468 (34%)	787 (31%)	336 (24%)	680 (27%)

TABLE A.3: Descriptive analysis for the pregnancy/infancy variables before and after imputation.

Exposures	Original Dataset N=4246	Imputed Dataset 4246*10
Sociodemographic and Psychosocial factors		
Age mean (sd) (years)	30.00 (5.10)	30.0 (5.11)
Education, n (%)		
Primary	1686 (42%)	12934 (30%)
Secondary	1106 (28%)	17788 (42%)
Tertiary	1218 (30%)	11738 (28%)
Working Condition, n (%)		
Paid job	3313 (79%)	33577 (79%)
No paid job activity	877 (21%)	8883 (21%)
Marital Status, n (%)		
Married or cohabiting	4046 (96%)	40568 (96%)
No partner	189 (4%)	1,892 (4%)
Household Income, n (%)		
Low	1233 (33%)	14,393 (34%)
Middle	1164 (31%)	13,114 (31%)
High	1339 (36%)	14,953 (35%)
Parity, n (%)		
Nullipara	2438 (58%)	24,705 (58%)
Multipara	1743 (42%)	17,755 (42%)
Baby's Sex, n (%)		
Boy	2181 (51%)	21,810 (51%)
Girl	2065 (49%)	20,650 (49%)
Anthropometric factors		
Pre-conception BMI, n (%)		
Underweight/normal	2716 (69%)	29,178 (69%)
Overweight/obese	1226 (31%)	13,282 (31%)
Gestational weight gain, n (%)		
0	894 (24%)	10,186 (24%)
1	1433 (38%)	16,210 (38%)
2	1417 (38%)	16,064 (38%)
Baby's Weight mean (sd) (g)	3203 (484.84)	3203 (484.53)
Baby's Length mean (sd) (cm)	48.83 (2.31)	48.82 (2.33)
BMI 6 months mean (sd)	17.27 (1.48)	17.27 (1.49)
BMI 1 year mean (sd)	17.56 (1.47)	17.56 (1.47)
BMI 2 years mean (sd)	16.65 (1.43)	16.68 (1.43)
Lifestyle behaviors		
Smoking Habits during Pregnancy, n (%)		
Never smoked	3391 (81%)	34,391 (81%)
Smoked	792 (19%)	8,069 (19%)
Breastfeeding, n (%)		
Never	1034 (26%)	11,286 (27%)
0-2.9 months	757 (19%)	8,186 (19%)
3-5.9 months	721 (18%)	7,755 (18%)
$\geq$6 months	1412 (36%)	15,233 (36%)
Age of first solid food, n (%)		
$<$4 months	262 (7%)	3,133 (7%)
4-5 months	2251 (62%)	26,313 (62%)
$\geq$6 months	1103 (31%)	13,014 (31%)
First solid food, n (%)		
Cereal porridge	2163 (58%)	24,566 (58%)
Fruit	213 (6%)	2,458 (6%)
Soup	1345 (36%)	15,436 (36%)
Clinical Exposures		
Hypertensive complications, n (%)		
Yes	408 (10%)	4,116 (9.7%)
No	3814 (90%)	38,344 (90%)
Gestational Diabetes, n (%)		
Yes	293 (7%)	2,956 (7.0%)
No	3929 (93%)	39,504 (93%)
Vaginal Delivery, n (%)		
Yes	2581 (62%)	26,120 (62%)
No	1553 (38%)	16,340 (38%)
Cesarean delivery, n (%)		
Yes	1612 (40%)	16,430 (39%)
No	2467 (60%)	26,030 (61%)
Gestational weeks mean (sd)	39.46 (1.62)	39.45 (1.63)

TABLE A.4: Descriptive analysis for the 4-year variables before and after imputation.

Exposures	**Original 4-year Dataset** N=4246	**Imputed 4-year Dataset** N=10*4130
Sociodemographic and Psychosocial		
Education, n%		
Primary	1555 (40%)	16,415 (40%)
Secondary	1116 (28%)	11,573 (28%)
Tertiary	1260 (32%)	13,312 (32%)
Working Condition, $n(\%)$		
Paid job	3312 (80%)	33,009 (80%)
No paid job activity	829 (20%)	8,291 (20%)
Marital Status, $n(\%)$		
Married or cohabiting	3715 (91%)	37,488 (91%)
No partner	378 (9%)	3,812 (9%)
Household Income, $n(\%)$		
Low	984 (24%)	10,036 (24%)
Middle	1185 (29%)	12,018 (29%)
High	1901 (47%)	19,246 (46%)
Sex		
Boy	2181 (51%)	21,150 (51%)
Girl	2065 (49%)	20,150 (49%)
Anthropometric Measures		
Z-score BMI, $n(\%)$		
Normal	2538 (68%)	28,106 (68%)
Overweight/obese	1194 (32%)	13,194 (32%)
Lifestyle behaviors		
Sedentary time median (min-max) (min/day)	92 (60-137)	91 (60, 137)
Active play median (min-max) (min/day)	120 (77-184)	120 (77, 180)
Hours of sports activity median (min-max) (min/week)	60 (0-120)	60 (0, 120)
Sleep duration, mean (sd) (hours)	10.10 (0.85)	10.00 (9.50, 10.50)
Fruit/Vegetables/Soup, n (%)		
$<$ 5 a day	1527 (43%)	17,688 (43%)
$\geq$ 5 a day	2052 (57%)	23,612 (57%)
Calories consumption median (min-max) (kcals)	1557 (1375-1742)	1,559 (1,377, 1,755)

TABLE A.5: Descriptive analysis for the 7-year variables before and after imputation.

Exposures	**Original 7-year Dataset** N=4246	**Imputed 7-year Dataset** N=10*3936
Sociodemographic and Psychosocial		
Education, $n(\%)$		
Primary	1394 (36%)	14,116 (36%)
Secondary	1172 (30%)	12,040 (31%)
Tertiary	1280 (33%)	13,204 (34%)
Working Condition, $n(\%)$		
Paid job	3233 (80%)	31,451 (80%)
No paid job activity	822 (20%)	7,909 (20%)
Marital Status, $n(\%)$		
Married or cohabiting	3562 (88%)	34,629 (88%)
No partner	497 (12%)	4,731 (12%)
Household Income, $n(\%)$		
Low	1062 (26%)	10,131 (26%)
Middle	1149 (28%)	11,120 (28%)
High	1842 (45%)	18,109 (46%)
Sex		
Boy	2181 (51%)	20,370 (52%)
Girl	2065 (49%)	18,990 (48%)
Anthropometric Measures		
Z-score BMI, $n(\%)$		
Normal	2400 (61%)	23,993 (61%)
Overweight/obese	1535 (39%)	15,367 (39%)
Mother's BMI, $n(\%)$		
Underweight/normal	1685 (45%)	17,774 (45%)
Overweight/obese	2031 (55%)	21,586 (55%)
Lifestyle behaviors		
Sedentary time median (min-max) (min/day)	161 (124-210)	161 (124, 210)
Active play median (min-max) (min/day)	73 (39-107)	73 (39, 107)
Hours of sports activity median (min-max) (min/week)	45 (0-150)	45 (0, 150)
Fruit/Vegetables/Soup, n (%)		
$<$ 5 a day	3572 (92%)	36,364 (92%)
$\geq$ 5 a day	294 (8%)	2,996 (8%)
Calories consumption median (min-max) (kcals)	1691 (1486-1908)	1,691 (1,485, 1,908)
Soft drinks, n (%)		
never or less than 1 per week	729 (19%)	7,305 (19%)
1-6 per week	1942 (49%)	19,459 (49%)
one or more per day	1255 (32%)	12,596 (32%)

TABLE A.6: Descriptive analysis for the 10-year variables before and after imputation.

Exposures	**Original 10-year Dataset** N=4246	**Imputed 10-year Dataset** N=10*4126
Sociodemographic and Psychosocial		
Education, $n(\%)$		
Primary	1336 (36%)	14,850 (36%)
Secondary	1139 (30%)	12,576 (30%)
Tertiary	1269 (34%)	13,914 (34%)
Household Income, $n(\%)$		
Low	1009 (25%)	10,568 (26%)
Middle	1154 (29%)	12,012 (29%)
High	1808 (46%)	18,760 (45%)
Sex		
Boy	2181 (51%)	21,260 (51%)
Girl	2065 (49%)	20,080 (49%)
Anthropometric Measures		
Z-score BMI, $n(\%)$		
Normal	2242 (57%)	23,452 (57%)
Overweight/obese	1703 (43%)	17,888 (43%)
Lifestyle behaviors		
Sedentary time		
median (min-max) (min/day)	154 (116-197)	154 (116, 197)
Hours of sports activity		
median (min-max) (min/week)	120 (0-240)	120 (0, 240)
Fruit/Vegetables/Soup, n (%)		
$<$ 5 a day	3156 (81%)	33,439 (81%)
$\geq$ 5 a day	749 (19%)	7,901 (19%)
Calories consumption		
median (min-max) (kcals)	1742 (1503-2058)	1,743 (1,504, 2,059)
Soft drinks, n (%)		
never or less than 1 per week	7723 (18%)	7,572 (18%)
1-6 per week	2183 (56%)	22,985 (56%)
one or more per day	1016 (26%)	10,783 (26%)

TABLE A.7: Main Effects of Each Variable on the Response with Corresponding Odds Ratios and p-Values

Exposures	Odds-Ratio	Adjusted p-Value
Age	0.98	0.15
Education		
Primary	1.74	< 0.01
Secondary	1.34	0.12
Tertiary	Ref	
Working Condition		
Paid job	Ref	
No paid job activity	1.32	0.05
Marital Status		
Married or cohabiting	Ref	
No partner	1.53	0.22
Household Income		
Low	Ref	
Middle	0.84	0.83
High	0.63	< 0.01
Parity		
Nullipara	Ref	
Multipara	1.12	1.00
Smoking Habits during Pregnancy		
Never smoked	Ref	
Smoked	1.35	0.03
Hypertensive complications		
Yes	1.62	< 0.01
No	Ref	
Gestational Diabetes		
Yes	1.67	0.01
No	Ref	
Vaginal Delivery		
Yes	0.83	0.27
No	Ref	
Cesarean delivery		
Yes	1.20	0.27
No	Ref	
Pre-conception BMI		
Underweight/normal	Ref	
Overweight/obese	2.48	< 0.01
Gestational weight gain		
Insufficient	1.04	1.00
Normal	Ref	
Excessive	1.60	< 0.01
Gestational weeks	1.00	1.00
Baby's Sex		
Boy	0.98	1.00
Girl	Ref	
Baby's Weight	1.17	0.54
Baby's Length	1.09	1.00
Breastfeeding		
Never	1.29	0.20
0-2.9 months	0.95	1.00
3-5.9 months	0.89	1.00
$\geq$ 6 months	Ref	
First solid food		
Cereal porridge	0.75	0.02
Fruit	0.81	1.00
Soup	Ref	
Age of first solid food		
$<$ 4 months	1.34	0.77
4-5 months	Ref	
$\geq$ 6 months	1.14	1.00
BMI 6 months	1.25	< 0.01
BMI 1 year	1.28	< 0.01
BMI 2 years	1.51	< 0.01

TABLE A.8: Main Effects of Each Variable on the Response with Corresponding Odds Ratios and p-values for the 4-year dataset.

Exposures	**Odds-Ratio**	**Adjusted p-Value**
Education		
Primary	1.78	< 0.01
Secondary	1.52	< 0.01
Tertiary	Ref	Ref
Working Condition		
Paid job	Ref	Ref
No paid job activity	1.17	0.60
Marital Status		
Married or cohabiting	Ref	Ref
No partner	1.22	0.69
Household Income		
Low	Ref	Ref
Middle	0.70	0.01
High	0.61	< 0.01
Z-score BMI		
Normal	Ref	Ref
Overweight/obese	6.04	< 0.01
Sedentary time	1.12	0.05
Active play	1.03	0.73
Hours of sports activity	1.05	0.69
Sleep duration	0.98	0.73
Fruit/Vegetables/Soup		
< 5 a day	Ref	Ref
≥ 5 a day	0.82	0.18
Calories consumption	1.13	0.73

TABLE A.9: Main Effects of Each Variable on the Response with Corresponding Odds Ratios and p-values for the 7-year dataset.

Exposures	**Odds-Ratio**	**Adjusted p-Value**
Education		
Primary	1.88	< 0.01
Secondary	1.57	< 0.01
Tertiary	Ref	Ref
Working Condition		
Paid job	Ref	Ref
No paid job activity	1.36	0.02
Marital Status		
Married or cohabiting	Ref	Ref
No partner	1.14	0.69
Household Income		
Low	Ref	Ref
Middle	0.79	0.16
High	0.62	< 0.01
Z-score BMI		
Normal	Ref	Ref
Overweight/obese	14.45	< 0.01
Maternal BMI		
Normal	Ref	Ref
Overweight/obese	2.67	< 0.01
Sedentary time	1.12	0.02
Active play	0.97	0.69
Hours of sports activity	0.95	0.16
Fruit/Vegetables/Soup		
< 5 a day	Ref	Ref
≥ 5 a day	1.41	0.16
Calories consumption	1.00	0.69
Soft drinks		
never or less than 1 per week	Ref	Ref
1-6 per week	1.16	0.69
one or more per day	1.37	0.05

TABLE A.10: Main Effects of Each Variable on the Response with Corresponding Odds Ratios and p-values for the 10-year dataset.

Exposures	**Odds-Ratio**	**Adjusted p-Value**
Education		
Primary	1.80	< 0.01
Secondary	1.49	< 0.01
Tertiary	Ref	Ref
Household Income		
Low	Ref	Ref
Middle	0.77	0.02
High	0.58	< 0.01
Z-score BMI		
Normal	Ref	Ref
Overweight/obese	28.42	< 0.01
Sedentary time	1.09	0.01
Hours of sports activity	0.93	< 0.01
Fruit/Vegetables/Soup		
< 5 a day	Ref	Ref
≥ 5 a day	0.88	0.21
Calories consumption	0.89	0.14
Soft drinks		
never or less than 1 per week	Ref	Ref
1-6 per week	1.29	0.02
one or more per day	0.99	0.91

TABLE A.11: Performance measures for each model.

Models	**Sensitivity**	**Specificity**	**AUC**	**PPA**	**NPA**	**Prediction Error**
4-Year Logistic Regression Model	0.61	0.87	0.76	0.71	0.80	23%
4-Year Model KNN	0.23	0.78	0.50	0.37	0.65	41 %
4-Year Model LDA	0.62	0.87	0.76	0.72	0.80	22%
4-Year Model Decision Tree	0.62	0.87	-	0.72	0.81	22 %
7-Year Logistic Regression Model	0.67	0.86	0.82	0.70	0.84	20%
7-Year Model KNN	0.53	0.49	0.51	0.34	0.67	50%
7-Year Model LDA	0.66	0.86	0.81	0.70	0.84	21%
7-Year Model Decision Tree	0.75	0.81	-	0.66	0.87	21 %
10-Year Logistic Regression Model	0.81	0.88	0.87	0.70	0.93	17 %
10-Year Model KNN	0.51	0.42	0.55	0.30	0.63	55%
10-Year Model LDA	0.81	0.88	0.88	0.70	0.93	17%
10-Year Model Decision Tree	0.81	0.88	0.88	0.70	0.93	17%

Bibliography

[1] A. Laxmaiah, M. Soric, J. J. Miranda, J. Bentham, P. Bovet, G. A. Stevens, A.-P. Kengne, M. Di Cesare, and Z. Bhutta, "The epidemiological burden of obesity in childhood: a worldwide epidemic requiring urgent action." 2019. [Cited on page 1.]

[2] "Organization WH obesity," https://www.who.int/health-topics/obesity#tab=tab_1, accessed: 2023-06-08. [Cited on page 1.]

[3] K. Sahoo, B. Sahoo, A. K. Choudhury, N. Y. Sofi, R. Kumar, and A. S. Bhadoria, "Childhood obesity: causes and consequences," *Journal of family medicine and primary care*, vol. 4, no. 2, p. 187, 2015. [Cited on pages 1 and 2.]

[4] E. P. Williams, M. Mesidor, K. Winters, P. M. Dubbert, and S. B. Wyatt, "Overweight and obesity: prevalence, consequences, and causes of a growing public health problem," *Current obesity reports*, vol. 4, no. 3, pp. 363–370, 2015. [Cited on page 1.]

[5] D. R. Thompson, E. Obarzanek, D. L. Franko, B. A. Barton, J. Morrison, F. M. Biro, S. R. Daniels, and R. H. Striegel-Moore, "Childhood overweight and cardiovascular disease risk factors: the national heart, lung, and blood institute growth and health study," *The Journal of pediatrics*, vol. 150, no. 1, pp. 18–25, 2007. [Cited on page 1.]

[6] "Organization WH obesity: Health consequences of being overweight 2013," https://www.who.int/news-room/questions-and-answers/item/obesity-health-consequences-of-being-overweight, accessed: 2023-06-08. [Cited on page 1.]

[7] M. Simmonds, A. Llewellyn, C. G. Owen, and N. Woolacott, "Predicting adult obesity from childhood obesity: a systematic review and meta-analysis," *Obesity reviews*, vol. 17, no. 2, pp. 95–107, 2016. [Cited on page 1.]

[8] S. Vuik, A. Lerouge, Y. Guillemette, A. Feigl, and A. Aldea, "The economic burden of obesity," 2019. [Cited on page 2.]

[9] E. Y. Lee and K.-H. Yoon, "Epidemic obesity in children and adolescents: risk factors and prevention," *Frontiers of medicine*, vol. 12, pp. 658–666, 2018. [Cited on page 2.]

[10] M. L. Endalifer and G. Diress, "Epidemiology, predisposing factors, biomarkers, and prevention mechanism of obesity: a systematic review," *Journal of obesity*, vol. 2020, 2020. [Cited on page 2.]

[11] C. P. Wild, "Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology," *Cancer Epidemiology Biomarkers & Prevention*, vol. 14, no. 8, pp. 1847–1850, 2005. [Cited on page 2.]

[12] ——, "The exposome: from concept to utility," *International journal of epidemiology*, vol. 41, no. 1, pp. 24–32, 2012. [Cited on page 2.]

[13] S. Santos, L. Maitre, C. Warembourg, L. Agier, L. Richiardi, X. Basagaña, and M. Vrijheid, "Applying the exposome concept in birth cohort research: a review of statistical approaches," *European journal of epidemiology*, vol. 35, pp. 193–204, 2020. [Cited on pages 2 and 13.]

[14] G. Ferrante, S. Fasola, G. Cilluffo, G. Piacentini, G. Viegi, and S. La Grutta, "Addressing exposome: An innovative approach to environmental determinants in pediatric respiratory health," *Frontiers in Public Health*, vol. 10, 2022. [Cited on page 2.]

[15] M. Vrijheid, S. Fossati, L. Maitre, S. Márquez, T. Roumeliotaki, L. Agier, S. Andrusaityte, S. Cadiou, M. Casas, M. de Castro *et al.*, "Early-life environmental exposures and childhood obesity: an exposome-wide approach," *Environmental Health Perspectives*, vol. 128, no. 6, p. 067009, 2020. [Cited on pages 2 and 13.]

[16] J. A. Kerr, C. Long, S. A. Clifford, J. Muller, A. N. Gillespie, S. Donath, and M. Wake, "Early-life exposures predicting onset and resolution of childhood overweight or obesity," *Archives of Disease in Childhood*, vol. 102, no. 10, pp. 915–922, 2017. [Cited on page 3.]

[17] G. Mascherini, C. Petri, E. Ermini, V. Bini, P. Calà, G. Galanti, and P. A. Modesti, "Overweight in young athletes: new predictive model of overfat condition," *International journal of environmental research and public health*, vol. 16, no. 24, p. 5128, 2019. [Cited on page 3.]

[18] N. Ziauddeen, S. Wilding, P. J. Roderick, N. S. Macklon, D. Smith, D. Chase, and N. A. Alwan, "Predicting the risk of childhood overweight and obesity at 4–5 years using population-level pregnancy and early-life healthcare data," *BMC medicine*, vol. 18, no. 1, pp. 1–15, 2020. [Cited on page 3.]

[19] M. Welten, A. H. Wijga, M. Hamoen, U. Gehring, G. H. Koppelman, J. W. Twisk, H. Raat, M. W. Heymans, and M. L. de Kroon, "Dynamic prediction model to identify young children at high risk of future overweight: Development and internal validation in a cohort study," *Pediatric obesity*, vol. 15, no. 9, p. e12647, 2020. [Cited on pages 3 and 13.]

[20] Y. Chen, C. Cai, J. Tan, X. Lei, Q. Chen, J. Zhang, and Y. Zhang, "High-risk growth trajectory related to childhood overweight/obesity and its predictive model at birth," *The Journal of Clinical Endocrinology & Metabolism*, vol. 107, no. 10, pp. e4015–e4026, 2022. [Cited on page 3.]

[21] P. S. Larsen, M. Kamper-Jørgensen, A. Adamson, H. Barros, J. P. Bonde, S. Brescianini, S. Brophy, M. Casas, G. Devereux, M. Eggesbø *et al.*, "Pregnancy and birth cohort resources in europe: a large opportunity for aetiological child health research," *Paediatric and perinatal epidemiology*, vol. 27, no. 4, pp. 393–414, 2013. [Cited on pages 3 and 35.]

[22] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793. [Cited on pages 4, 7, 8, 9, and 12.]

[23] C. K. Enders, *Applied missing data analysis*. Guilford Publications, 2022. [Cited on page 4.]

[24] G. Kauermann, H. Küchenhoff, and C. Heumann, *Statistical Foundations, Reasoning and Inference*. Springer, 2021. [Cited on pages 4, 8, 10, and 11.]

[25] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, no. 1, pp. 1–17, 2013. [Cited on pages 4, 9, 17, and 18.]

[26] S. Meeyai, "Logistic regression with missing data: a comparisson of handling methods, and effects of percent missing values," *Journal of Traffic and Logistics Engineering*, vol. 4, no. 2, 2016. [Cited on pages 4, 10, 12, and 17.]

[27] K. J. Lee and J. B. Carlin, "Recovery of information from multiple imputation: a simulation study," *Emerging themes in epidemiology*, vol. 9, no. 1, pp. 1–10, 2012. [Cited on pages 4 and 17.]

[28] 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 4, 13, and 20.]

[29] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014. [Cited on pages 4, 23, 24, and 26.]

[30] L. Freijeiro-González, M. Febrero-Bande, and W. González-Manteiga, "A critical review of lasso and its derivatives for variable selection under dependence among covariates," *International Statistical Review*, vol. 90, no. 1, pp. 118–145, 2022. [Cited on page 4.]

[31] M. Welten, M. L. de Kroon, C. M. Renders, E. W. Steyerberg, H. Raat, J. W. Twisk, and M. W. Heymans, "Repeatedly measured predictors: a comparison of methods for prediction modeling," *Diagnostic and prognostic research*, vol. 2, no. 1, pp. 1–10, 2018. [Cited on pages 4 and 31.]

[32] V. Shetty, C. H. Morrell, and S. S. Najjar, "Modeling a cross-sectional response variable with longitudinal predictors: an example of pulse pressure and pulse wave velocity," *Journal of applied statistics*, vol. 36, no. 6, pp. 611–619, 2009. [Cited on pages 4 and 31.]

[33] Y.-H. Chen, K. K. Ferguson, J. D. Meeker, T. F. McElrath, and B. Mukherjee, "Statistical methods for modeling repeated measures of maternal environmental exposure biomarkers during pregnancy in association with preterm birth," *Environmental Health*, vol. 14, no. 1, pp. 1–13, 2015. [Cited on pages 4 and 31.]

[34] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976. [Cited on pages 7 and 12.]

[35] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *Journal of school psychology*, vol. 48, no. 1, pp. 5–37, 2010. [Cited on pages 9, 10, and 12.]

[36] C. M. Musil, C. B. Warner, P. K. Yobas, and S. L. Jones, "A comparison of imputation techniques for handling missing data," *Western journal of nursing research*, vol. 24, no. 7, pp. 815–829, 2002. [Cited on page 11.]

[37] P. D. Allison *et al.*, *Missing data*. Sage Thousand Oaks, CA, 2010, vol. 200210, no. 9781412985079.31. [Cited on page 11.]

[38] G. Molenberghs and M. Kenward, *Missing data in clinical studies*. John Wiley & Sons, 2007. [Cited on pages 12 and 18.]

[39] D. B. Rubin, *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004, vol. 81. [Cited on pages 12, 16, 17, and 18.]

[40] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 1997. [Cited on page 12.]

[41] S. Van Buuren, J. P. Brand, C. G. Groothuis-Oudshoorn, and D. B. Rubin, "Fully conditional specification in multivariate imputation," *Journal of statistical computation and simulation*, vol. 76, no. 12, pp. 1049–1064, 2006. [Cited on page 13.]

[42] S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2018. [Cited on pages 13, 14, and 15.]

[43] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, P. Solenberger *et al.*, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, vol. 27, no. 1, pp. 85–96, 2001. [Cited on page 13.]

[44] C. Warembourg, L. Maitre, I. Tamayo-Uria, S. Fossati, T. Roumeliotaki, G. M. Aasvang, S. Andrusaityte, M. Casas, E. Cequier, L. Chatzi *et al.*, "Early-life environmental exposures and blood pressure in children," *Journal of the American College of Cardiology*, vol. 74, no. 10, pp. 1317–1328, 2019. [Cited on page 13.]

[45] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011. [Cited on page 13.]

[46] 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 13.]

[47] 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 16.]

[48] A. Marshall, D. Altman, R. Holder, and P. Royston, "Combining estimates of interest in prognostic 498 modelling studies after multiple imputation: Current practice and guidelines," *BMC Med Res 499 Methodol*, vol. 9, no. 57, p. 500, 2009. [Cited on page 16.]

[49] J. Barnard and D. B. Rubin, "Miscellanea. small-sample degrees of freedom with multiple imputation," *Biometrika*, vol. 86, no. 4, pp. 948–955, 1999. [Cited on page 17.]

[50] J. H. Lee, J. Huber Jr *et al.*, "Multiple imputation with large proportions of missing data: How much is too much?" in *United Kingdom Stata Users' Group Meetings 2011*, no. 23. Stata Users Group, 2011. [Cited on page 17.]

[51] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, "The proportion of missing data should not be used to guide decisions on multiple imputation," *Journal of clinical epidemiology*, vol. 110, pp. 63–73, 2019. [Cited on page 18.]

[52] P. Ranganathan, C. Pramesh, R. Aggarwal *et al.*, "Common pitfalls in statistical analysis: Measures of agreement," *Perspectives in clinical research*, vol. 8, no. 4, p. 187, 2017. [Cited on pages xi, 18, and 19.]

[53] L. Daly and G. J. Bourke, *Interpretation and uses of medical statistics*. John Wiley & Sons, 2008. [Cited on pages 18 and 19.]

[54] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971. [Cited on page 19.]

[55] S. van Buuren, K. Groothuis-Oudshoorn, A. Robitzsch, G. Vink, L. Doove, S. Jolani *et al.*, "Package 'mice'," *Computer software*, 2015. [Cited on page 20.]

[56] J. Honaker, G. King, and M. Blackwell, "Amelia ii: A program for missing data," *Journal of statistical software*, vol. 45, pp. 1–47, 2011. [Cited on pages 20 and 21.]

[57] D. J. Stekhoven and M. D. J. Stekhoven, "Package 'missforest'," *R package version*, vol. 1, 2013. [Cited on page 21.]

[58] Y.-S. Su, A. Gelman, J. Hill, and M. Yajima, "Multiple imputation with diagnostics (mi) in r: Opening windows into the black box," *Journal of Statistical Software*, vol. 45, pp. 1–31, 2011. [Cited on page 21.]

[59] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of statistical software*, vol. 36, pp. 1–13, 2010. [Cited on page 25.]

[60] M. J. Hasan, J. Kim, C. H. Kim, and J.-M. Kim, "Health state classification of a spherical tank using a hybrid bag of features and k-nearest neighbor," *Applied Sciences*, vol. 10, no. 7, p. 2525, 2020. [Cited on pages ix, 25, and 26.]

[61] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Cited on page 26.]

[62] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006. [Cited on page 26.]

[63] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. [Cited on page 26.]

[64] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005. [Cited on page 26.]

[65] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *Annals of statistics*, vol. 37, no. 4, p. 1733, 2009. [Cited on page 26.]

[66] A. M. Wood, I. R. White, and P. Royston, "How should variable selection be performed with multiply imputed data?" *Statistics in medicine*, vol. 27, no. 17, pp. 3227–3246, 2008. [Cited on page 28.]

[67] J. Du, J. Boss, P. Han, L. J. Beesley, M. Kleinsasser, S. A. Goutman, S. Batterman, E. L. Feldman, and B. Mukherjee, "Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods," *Journal of Computational and Graphical Statistics*, vol. 31, no. 4, pp. 1063–1075, 2022. [Cited on pages 29 and 30.]

[68] Q. Chen and S. Wang, "Variable selection for multiply-imputed data with application to dioxin exposure study," *Statistics in medicine*, vol. 32, no. 21, pp. 3646–3659, 2013. [Cited on page 30.]

[69] B. Everitt, *Finite mixture distributions*. Springer Science & Business Media, 2013. [Cited on page 33.]

[70] F. Leisch, "Flexmix: A general framework for finite mixture models and latent glass regression in r," 2004. [Cited on pages 33 and 34.]

[71] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression," *Journal of classification*, vol. 5, pp. 249–282, 1988. [Cited on page 33.]

[72] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977. [Cited on page 34.]

[73] B. Gruen, F. Leisch, D. Sarkar, F. Mortier, N. Picard, and M. B. Gruen, "Package 'flexmix'," 2015. [Cited on page 34.]

[74] G. A. of the World Medical Association *et al.*, "World medical association declaration of helsinki: ethical principles for medical research involving human subjects," *The Journal of the American College of Dentists*, vol. 81, no. 3, pp. 14–18, 2014. [Cited on page 35.]

[75] M. d. Onis, A. W. Onyango, E. Borghi, A. Siyam, C. Nishida, and J. Siekmann, "Development of a who growth reference for school-aged children and adolescents," *Bulletin of the World health Organization*, vol. 85, no. 9, pp. 660–667, 2007. [Cited on pages 37 and 38.]

[76] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *Statistics in medicine*, vol. 38, no. 11, pp. 2074–2102, 2019. [Cited on page 54.]

[77] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979. [Cited on page 57.]