U.PORTO

MASTER
ECONOMICS OF BUSINESS AND STRATEGY

Application of longitudinal models for evaluating cohort and age effects on anthropometric measurements of children from Porto – report of an internship at ISPUP

Carolina Tavares de Abreu Ramos de Oliveira

M

2023

U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Application of longitudinal models for evaluating cohort and age effects on anthropometric measurements of children from Porto – report of an internship at ISPUP

Carolina Tavares de Abreu Ramos de Oliveira

Internship report
Master in Economics of Business and Strategy

Supervised by
Prof. Doutor Álvaro Fernando Santos Almeida
Prof. Doutor Milton Severo Barros da Silva

2023

SCHOOL OF
ECONOMICS
AND MANAGEMENT
UNIVERSITY OF PORTO

Application of longitudinal models for evaluating cohort and age effects on anthropometric measurements of children from Porto – report of an internship at ISPUP

03.07.2023 | Carolina Tavares de Abreu Ramos de Oliveira

ACKNOWLEDGMENTS

The person I am today has a little bit of each person who was part of my path, academic and personal wise. Therefore, I could not start this report except by expressing the inexplicable gratitude I feel for the people around me and for the life I have been leaving.

First of all, I would like to start by thanking my supervisor at ISPUP, Prof. Doctor Milton Severo who despite his always busy schedule, always found a little hole for me and guided and supported all my work in the last few months. I would like to thank him for his flexibility and understanding throughout the entire process.

To my supervisor at FEP, Prof. Doctor Álvaro Almeida, for the suggestions, availability and promptness in helping, clarifying and guiding each step of this journey, especially in writing this report.

To my friends from the heart and of all hours: nothing would be the same without you. Thank you for for encouraging me all the time and always celebrate my achievements as if they were your own. Friends really are the family we choose and you make me feel like I chose very well.

To my family, that allowed me to fly and live my dreams, while still being the heaven to which I always wanted to return. Thank you for the unconditional love, support, for trusting me, for always reminding me of what really matters and always making me believe that I'm capable of anything.

I'm so grateful for everything. May this be the beginning of the end of this great chapter of my life.

ABSTRACT

This report is part of the Master's Degree in Economics of Business and Strategy at the Faculty of Economics of the University of Porto (FEP) and concerns an internship carried out at the Institute of Public Health of the University of Porto (ISPUP) that was both a curricular internship and an integrant one as part of the Quantitative Techniques for Economics and Management (QTEM) network, of which FEP is a partner university.

In this project, the Epidemiological Health Investigation of Teenagers (EPITeen) and Generation 21 (GXXI) cohorts of individuals, which have been longitudinally followed since 2003 and 2005, respectively, were considered. Using the data from these groups of participants, the primary objective of this study was to examine how habits and behaviours acquired during adolescence can influence adult health outcomes. To achieve this, two main analytical approaches were employed. Firstly, the cohort effect was explored to understand the differences between individuals born at different time periods, focusing on variables such as Body Mass Index (BMI), weight and height. Secondly, the age effect was examined to analyse the changing incidence and prevalence of a disease or health condition, such as obesity, as participants aged.

To address these objectives, some statistical techniques were used, such as the linear mixed-effects model, the predominant technique for examining longitudinal data. Additionally, regression splines were incorporated into the models to capture the non-linear behavior of variables over time and the Generalized Additive Models for Location, Scale and Shape (GAMLSS) model was employed to enhance the overall model fit.

Integrating the ISPUP statistics team, the internship focused on supporting the analytical part of the study, using data cleaning and visualization. By carefully examining the data and employing robust statistical techniques, this project aimed to provide evidence-based insights into the long-term implications of adolescent habits on their adult health. These findings can ultimately contribute to the development of effective interventions and policies to improve public health and well-being.

Key words: *Cohort, Longitudinal data, Mixed-effect models, Splines, GAMLSS, Public Health*

RESUMO

O presente relatório insere-se no âmbito do Mestrado em Economia da Empresa e da Estratégia da Faculdade de Economia da Universidade do Porto (FEP) e diz respeito a um estágio realizado no Instituto de Saúde Pública da Universidade do Porto (ISPUP), que foi simultaneamente um estágio curricular e um estágio integrado no âmbito da rede internacional *Quantitative Techniques for Economics and Management* (QTEM), da qual a FEP é universidade parceira.

Neste projeto, foram consideradas as coortes de indivíduos Epidemiological Health Investigation of Teenagers (EPITeen) e Geração 21 (GXXI), que têm sido seguidas longitudinalmente desde 2003 e 2005, respetivamente. Utilizando os dados destes grupos de participantes, o objetivo principal deste estudo foi analisar de que forma os hábitos e comportamentos adquiridos durante a adolescência podem influenciar o estado de saúde na idade adulta. Para atingir este objetivo, duas abordagens principais foram utilizadas. Por um lado, o efeito de coorte foi explorado para compreender as diferenças entre indivíduos nascidos em diferentes períodos de tempo, centrando-nos em variáveis como o Índice de Massa Corporal (IMC), o peso e a altura. Em segundo lugar, o efeito da idade foi examinado para analisar a alteração da incidência e prevalência de uma doença ou estado de saúde, como a obesidade, à medida que os participantes envelheciam.

Para dar resposta a estes objectivos, foram utilizadas várias técnicas estatísticas, como o modelo linear de efeitos mistos – a técnica predominante para a análise de dados longitudinais. Adicionalmente, foram incorporados splines de regressão nos modelos para captar o comportamento não linear das variáveis ao longo do tempo e o modelo aditivo generalizado para locação, escala e forma (GAMLSS) foi empregue para melhorar o ajuste global do modelo.

Tendo integrado a equipa de estatística do ISPUP, o trabalho centrou-se no apoio à parte analítica do estudo, utilizando a limpeza e visualização de dados. Examinando cuidadosamente os dados e empregando técnicas estatísticas robustas, este projeto teve como objetivo fornecer conclusões baseadas em evidências relativas às implicações a longo prazo dos hábitos dos adolescentes na sua saúde em adultos. Estas conclusões podem, em última análise, contribuir para o desenvolvimento de intervenções e políticas eficazes para melhorar a saúde pública e o bem-estar geral da população.

Palavras-chave: *Coorte, Dados Longitudinais, Lodelos Lineares de efeitos misto, Splines, GAMLSS, Saúde Pública*

INDEX

FIGURE'S INDEX

TABLE'S INDEX

ANNEXES' INDEX

1. INTRODUCTION

A disease or health condition, whatever it is and depending on its degree of dispersion and severity, generates benefits and costs in a society as the economy shapes and responds to its dispersion (Dangerfield C., 2022). Naturally, the definition is relatively shaky but epidemiology can be defined as the study of the determinants and distribution of health-related states or events, including disease, and the application of this knowledge to the control of disease and other health problems (Stamuli & Panagiotakos, 2022). Specifically in epidemiological analysis, it is crucial to understand how people's behaviour evolve over time, how it manifests in their health and which specific variables are important to be considered. Not everyone wins or loses, but it is certain that people adjust their behaviour so that this distribution of benefits vs costs is redistributed and moulded to optimise the general wellbeing of the community. Being aware of the costs associated with certain diseases and intervention in this field and delving into how those evolve using the right methods becomes essential to build a useful approach that integrates epidemiology and economics efficiently (Dangerfield, C., 2022).

Public health research plays a pivotal role in understanding the determinants of health outcomes and designing effective interventions. Longitudinal studies provide valuable insights into the long-term implications of various factors on individuals' health. In this context, the EPITeen and GXXI cohorts, conducted at ISPUP, have emerged as significant resources for investigating the impact of adolescent behaviors on adult health outcomes (Costa et al., 2018)

This report will explore the longitudinal data analysis that was conducted in this context, using linear mixed-effects models, a well-established approach for analyzing repeated measures data and accounting for within-subject correlation over time (Verbeke et al., 1997). Additionally, non-linear patterns in the data were captured by incorporating regression splines into the models, allowing for flexible modeling of the relationship between variables and their change over time (Wood, 2006). The use of GAMLSS further enhanced the modeling framework, enabling the examination of more complex relationships between predictors and health outcomes (Rigby & Stasinopoulos, 2005).

Quantitative techniques, especially statistical analysis, have proven instrumental in uncovering patterns, relationships, and trends within complex datasets. They enable researchers to extract meaningful insights from large-scale longitudinal studies and provide evidence-based guidance for public health interventions (Szklo et al., 2001). Statistical analysis allows for the exploration of associations between risk factors and health outcomes, helping to identify potential causal relationships and inform policy

decisions (Rothman et al., 2008). These topics involve sensitive aspects of people's lives and have a direct impact on their well-being directly, requiring a rigorous and meticulous approach, so the use of appropriate statistical tools ensures that the conclusions drawn from the analysis are precise and well-founded. As a result of the employment of robust statistical techniques and a careful examination of the data, the conclusions generated from this research can ultimately contribute to the development of effective interventions and policies to improve public health and well-being.

The present report is the result of a reflection process on the experiences lived, the lessons learned and the challenges faced during this internship. Its aim is to describe and analyse the main results obtained within this project, while also giving relevance to and exploring the theory behind it. In this sense, it is divided into five main chapters. The present and first chapter includes a generic contextualization regarding the relevance of the themes that this report will address, as well as highlighting the importance of investigating them using the appropriate statistical methods when you want to obtain reliable results. The second chapter will report more extensively on the details of the internship itself, the context in which it emerged and why was it so important in my academic path. It will also include a description of its main objectives, the tasks performed and the role assumed in the institution. Chapter 3 will present an extensive literature review on all relevant topics to the development of this project, namely linear mixed effects models, regressions splines, gamlss. Chapter 4 will present the project in which I carried out most of my internship tasks and which therefore allowed me to obtain several interesting conclusions and results. After explaining the origin of the data used and contextualizing the study carried out, this chapter will explore the various stages of data cleaning, treatment and description and, subsequently, the respective results obtained, specifically at the level of mixed effects and GAMLSS. The fifth and final chapter will conclude by summarizing all the main findings and contributions of the work, as well as some of its limitations.

2. THE INTERNSHIP

2.1. The host institution - Institute of Public Health of the University of Porto (ISPUP)

A non-profit association whose action plan focuses on research and the deepening of knowledge around several topics, all related to public health. This is how we can in a generalised way define what ISPUP is. Founded in 2006, this renowned research institution counts with group of human resources ranging from researchers to mathematicians to health professionals such as nutritionists, psychologists, etc whose common mission is to contribute to the generation and transmission of new and useful knowledge in this field. Their main objectives include conducting cutting-edge research, contributing to evidence-based policies and interventions that address public health challenges and providing high-quality training in this field. ISPUP is allied to the University of Porto, acting according to its motto that

"Research in public health saves lives"

In each project it is associated with, the institution's intervention is aimed at protecting the population, by creating and disseminating information that enables prevention. The type of research it carries out, specifically in its Epidemiology Research Unit (EPIUnit) has an innovative and highly scientific character, seeking to objectively provide solutions to global health problems. Their research activities span a wide range of areas, encompassing epidemiology, biostatistics, genetics, environmental health, social sciences, and health promotion. ISPUP's researchers investigate diverse topics, such as the determinants of health and disease, risk factors, health behaviours, healthcare access and utilization, and health inequalities. In this sense, the EPIUnit is organised based on a variety of research groups made up of personnel who guarantee a very high specialization in each of these areas, and who are distributed as indicated below in Figure 1.

In addition, ISPUP also coordinates the Multidisciplinary Unit of Biomedical Research of the Institute of Biomedical Sciences Abel Salazar of the University of Porto (UMIB) and the Centre for Research in Physical Activity, Health and Leisure of the Faculty of Sports of the University of Porto (CIAFEL) - all research and development units whose missions align with ISPUP's, to integrate public health research in its activities. However, their activity is not only related to autonomous research since ISPUP is also actively involved in training the next generation of public health professionals. The

institution provides support in the implementation of projects within the scope of post-graduate courses in this area and also promotes intensive educational programs, essentially for resident doctors with this specific training. In addition, it also organizes workshops, seminars and conferences to disseminate research findings and promote scientific collaboration.

Through all of these years of research and educational endeavors, ISPUP has collaborated with various national and international partners to foster interdisciplinary research, create knowledge that informs health policies and advances public health practices and ultimately creates a positive impact on public health and improves population's well-being.

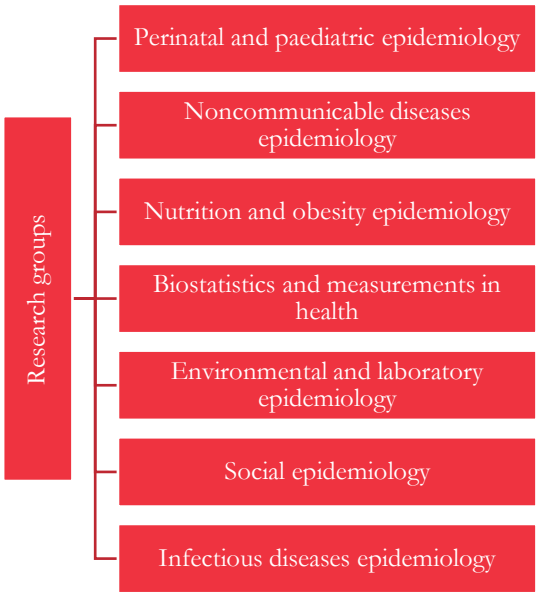


Figure 1: Research groups in the EPIUnit at ISPUP (ISPUP, 2021).

2.2. Internship contexto and description of the role

The present internship took place simultaneously as a curricular internship within the scope of the Master's degree in Economics of Business and Strategy at FEP and as an integrant internship in the Quantitative Techniques for Economics and Management (QTEM) network, of which FEP is a partner university.

In a brief contextualization, QTEM consists of a very technical program including two periods of exchange in two other universities that are also academic partners of the network and that provides a strong quantitative skills-building component - namely on Big Data Analytics, Decision Support and Management. Moreover, joining this network also allowed my participation in an international competition of business cases solved between multinational teams of QTEM students (Global Business Analytics Challenge), in which we were set the challenge of leveraging data science in recommendations to enhance the work of the British Red Cross. Also in this scope, I also had intensive trainings with two essential focuses: 1) Data Science - more technical skills in softwares like R, Python, etc - and 2) Digital Leadership - more oriented towards the development of soft skills. As part of this program, I was required to take a quantitative internship in which I could put into practice the skills acquired so far from the program, namely around the topics mentioned above - Quantitative Methods, Data Analytics, etc. This internship at ISPUP arises in this context as the perfect way to reconcile my various academic needs with a project that would challenge me technically and help me consolidate all the knowledge acquired so far, putting it into practice.

During the period of this internship, I had the opportunity to improve my skills and knowledge in the area of statistics, specifically biostatistics, and apply them in practice, working on projects and activities related to various topics that will be described in detail later in this report.

From 1 February to 31 May 2023, in a total of 512h, I was integrated in the biostatistics team of ISPUP, specifically in the EPIUnit, supporting highly experienced researchers and mathematicians, namely with a quantitative and statistical focus, which was the main objective of my internship and my duties. This group of researchers, specifically called "Biostatistics and Measurements in Health", develops and implements statistical and mathematical methods of analysis in public health research. Its focus is essentially at the level of 1) statistical modelling and 2) measurement of the current health status of the population.

Integrating this team, my work included collaboration in consultations carried out at ISPUP in the area of statistics. More specifically, the cleaning of databases of information and the descriptive analysis of that same information. Furthermore, I handled and trated large amounts of data in order to build longitudinal models within the information cohorts existing at ISPUP - Epiteen and Generation XXI – and to choose the best parameterization of the models in order to better respond to the analysis. Participation in scientific seminars and meetings of the statistics group were also part of the weekly tasks.

The calendar of activities developed during the internship, with a slightly more detailed description, can be found in Annex 1.

Besides, still within the scope of this internship, I also attended the course "Decision Support Systems in Health" taught by Prof. Dr. Milton Severo Barros da Silva, which allowed me to reinforce my specific knowledge in the following subjects: Cluster Analysis (hierarchical methods and non-hierarchical methods) and Classification Models like classification trees, Naive Bayes, Knn, etc.

All the quantitative work carried out was developed in R software. R is an open source software environment and programming language used in statistical analysis for data manipulation, data analysis and graphical representation and visualization. Offering a wide range of techniques, including linear and nonlinear regression, time series analysis, survival and multivariate analysis, among others, it is a super powerful tool for conducting sophisticated research and analysis in fields such as social sciences, economics, biology, etc. It also excels in statistical modelling and hypothesis testing, allowing to effectively explore relationships between data, identify patterns and draw informed conclusions with reproducible results (R Core Team, 2021).

3. LITERATURE REVIEW

3.1. Linear mixed-effects models

It is undeniable that linear models can provide to a wider understanding of many aspects of the world. Through those, people are able to evidence relationships in data in a powerful way (James et al., 2013). In a very simple approach, we can describe this idea as **Y ~ X**, meaning that Y is the dependent variable that **X**, the independent variables or exposures, will seek to predict or explain. Briefly, this type of model describes how data points collected regarding variable X - the predictors - are going to be related to the responses retrieved for variable Y. This relationship is described through a simple mathematical representation. Adding a bit of reality to this approach, it is understood that this prediction is systematically influenced by external and uncontrollable factors that cannot be ignored, so it is more correct to represent this idea as Y ~ **X** + ε , where ε represents this error character (Winter, 2013). Thus, we can consider that these models have one or more fixed effects, X's, and an error term, ε. Generically, this is represented as in 3.1.:

$$y = \beta_0 + \sum \beta_i X_i + \varepsilon_i \qquad \textbf{(3.1.)}$$

in which β represents linear parameter estimates to be estimated.

These models are called linear because a change in the independent variable leads to a linear change in the dependent variable, leading to a straight line representation when plotted on a graph. Still, for these models to be valid there are some conditions that must be met and testing whether the assumption of independent observations is met is one of the most important steps in building a consistent and reliable model (Winter, 2013). It is crucial to ensure the independence of the data used. In a linear analysis, if each element in the dataset comes from a different subject, then independence is guaranteed. However, if several outputs arise from each subject, the responses coming from the same subject will not be considered independent of each other and, therefore, we will have the opposite characteristic - interdependence of elements (Rencher & Schaalje, 2008). Unfortunately, it is not so rare that this condition is violated, constituting the risk of obtaining illogical results (Brown and Prescott, 2006).

People often want to conduct certain studies that involve collecting more data per individual, leading to interdependent measures. This can be a reflection of heterogeneity due to unaccountable factors

(Diggle et al., 2002). This is the case with the observations obtained for the project that is detailed in this report, where, in the data collected, each subject provides multiple observations for several variables, as this is a longitudinal analysis. To try to resolve this issue, random effects can be added for each subject, generating what are called linear mixed effects models, defining a different baseline value/intercept for each individual (Winter, 2013).The next parts will elaborate on this topic.

Linear mixed-effects models are often seen as a vehicle for longitudinal data analysis (Zhang, 2004). Longitudinal analysis allows us to explore dynamic rather than just static concepts, as we are considering a given number of measurements for a specific successive time, allowing us to explore their evolution over time (Twisk, 2013). Since this type of data tracks the same information about the same subjects, usually the same sample, at various points in time, the observations over time are correlated. For this reason, compliance with the condition of independence of observations is conditioned. What will mark the difference in these models in face of the normal linear models defined above is the mixture between fixed and random effects. In fact, this is exactly where the term "mixed" comes from. The fact that these models contain fixed effects (β) and random effects (b) allows them to model and quantify the relationship between a response variable and one or more explanatory variables, while accounting for the existing correlation among the observations within each subject. For this reason, those can also be referred to as random coefficient analysis or multilevel analysis (Laird and Ware, 1982). Here is an example of their appearance:

$$y_i = X_i'\beta + Z_i'\boldsymbol{b_i} + \varepsilon_i \qquad \textbf{(3.2.)}$$

in which X and Z are matrices of covariates, $\boldsymbol{\beta}$ represents the fixed effects and $\boldsymbol{b}$ represents the random effects, representing the variability among individuals, and assuming a normal distribution. ε_i also represents intra-individual variability.

The main idea regarding this analysis is that the regression coefficients can vary between subjects. What happens is that, to the notation Y ~ X + ε used before, we now add Y ~ X + (1|subject) + ε, meaning that we must consider a different intercept for each subject, where 1 represents the intercept. This approach will provide the analysis with great flexibility because of the wide variety of ways in which different observations can be correlated with each other (Brown and Prescott, 2006). Basically, knowing that we will have several responses per subject, using this method we can indicate to the model that each response will depend on the baseline value previously defined for each subject. Therefore, it will assume

repeated observations for a specific subject to be independent (Diggle et al., 2002). This idea can however become more complex depending on what we want to vary. As an example, a simple illustration prepared in R is provided below, in which I used the data that will be used later in this analysis.

Naturally, in the simplest version of the model, we use a simple linear regression in which only the constant, b_{0i} , has a random effect and therefore varies. The effect over time is assumed to remain constant.

$$Y_{it} = \beta_0 + \beta_{1t} + \boldsymbol{b_{0i}} + \varepsilon_{it} \qquad \textbf{(3.3.)}$$

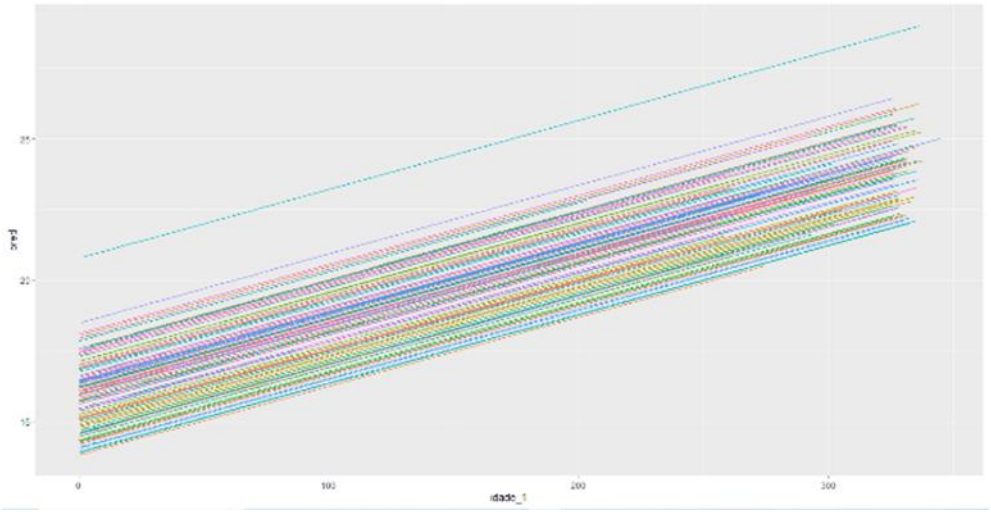


Figure 2: Development over time of a particular outcome variable Y - different intercepts for different subjects (Source: own elaboration).

In figure 3.1., Y_{it} represents the response variable for the t-th observation within the i-th subject; β_0 is the fixed intercept that represents the average response across all subjects; b_{0i} represents the random effect for the i-th subject as already mentioned and captures the deviation of the response for that subject from the overall average. This constant, b_{0i} , is assumed to follow a normal distribution with mean zero and a variance parameter σ^2; Again, ε_{it} is the residual error term (Bates et al., 2015). By including the random effect, this approach captures the variation in intercepts between subjects.

A slightly different approach is the model in which the effect over time varies between subjects, keeping the intercept constant. In this case, the random effect is associated with the slope and this can be

called a random slope model. Below there is the representation of this case in which the slope, $\boldsymbol{b_{1it}}$, is a random variable.

$$Y_{it} = \beta_0 + \beta_{1it} + \boldsymbol{b_{1it}} + \varepsilon_{it} \qquad \textbf{(3.4.)}$$

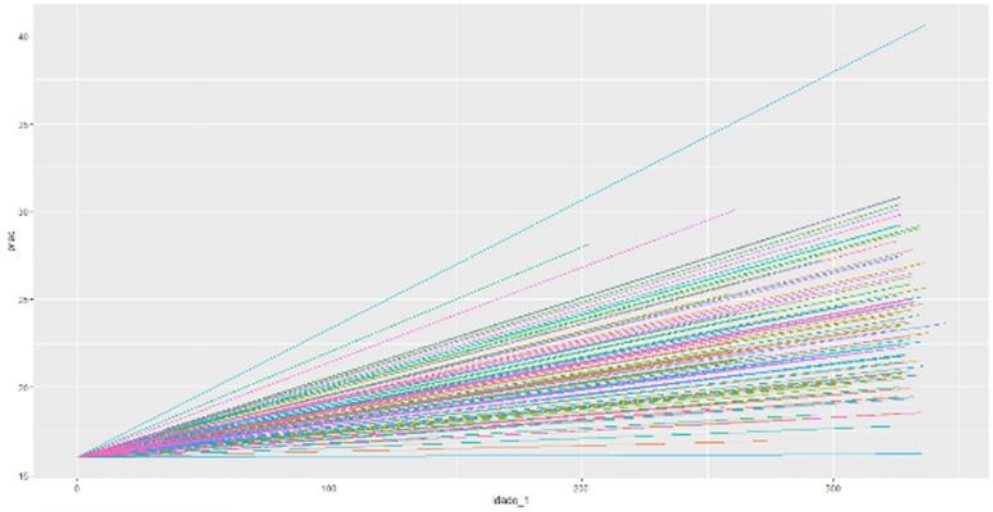


Figure 3: Development over time of a particular outcome variable Y - different slopes for different subjects (Source: own elaboration).

Finally, if we want to vary the intercept and the effect over time from subject to subject, both the constant, $\boldsymbol{b_{0i}}$, and the slope, $\boldsymbol{b_{1it}}$, will be random variables.

$$Y_{it} = \beta_0 + \beta_{1it} + \boldsymbol{b_{0i}} + \boldsymbol{b_{1it}} + \varepsilon_{it} \qquad \textbf{(3.5.)}$$

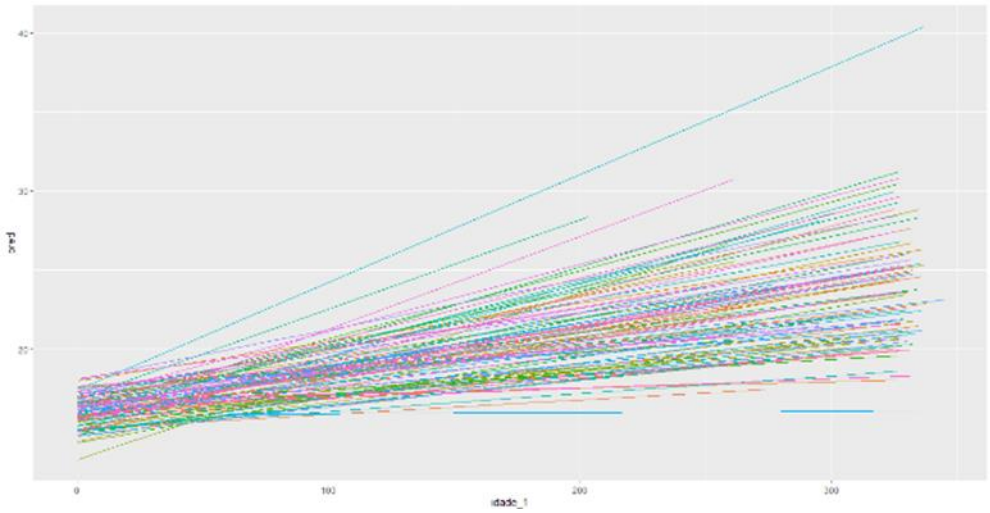


Figure 4: Development over time of a particular outcome variable Y - different intercepts and different slopes for different subjects (Source: own elaboration).

Despite being simple and intuitive to apply, in certain analyses, the goodness of fit that can be obtained from a linear model is not sufficient. This happens because the linearity they assume is not always exact. As already mentioned, some important conditions need to be met for linearity to exist and all the associated complexity means that, in reality, this characteristic is only approximately guaranteed, which may lead to results that are poor in precision and reliability (James et al., 2013). Thus, relaxing this linearity assumption may be useful to model the non-linear behaviour of the evolution of variables over time. Now deepening into some of its extensions.

Polynomial Regressions are a simple yet good example of an extension of the linear model that arises by introducing polynomial terms. As the proper name indicates, additional predictors are added to the ones included in the linear model, by raising each of its predictors to a power. This will give a non-linear fit to the data and therefore the fixed effects undergo a nonlinear transformation (Hastie et al., 2009). The relation between the dependent variable and the predictors is described through the following function

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 + \ldots + \beta_d {x_i}^d + \varepsilon_i \qquad \textbf{(3.6.)}$$

representing a non-linear link between the two. Here, the degree of the polynomial, d, represents the number of terms included in the equation. Thus, a quadratic regression makes the prediction using two variables , X and X^2; a cubic one uses three - X, X^2 and X^3 (Ostertagová, E., 2012). Incorporating these higher-order terms can add a lot of value to an analysis since it captures bends and curvatures that would not be captured with the linear regression (Hastie & Tibshirani, 1990). However, the increase of d can also result in a problem of overfitting if the model becomes too closely tailored to the data, resulting in bad results.

Moreover, there are other more advanced techniques that also constitute important extensions to linear models and whose analysis is very relevant in the scope of this report. Namely, in the following sections we go deeper into Regression Splines and GAMLSS.

3.2. Regression Splines

Regression splines are a statistical technique used for modeling nonlinear relationships between variables in regression analysis. They involve dividing the range of the independent variable into smaller

segments and fitting a polynomial function to each segment (James et al., 2013). The polynomial functions used are typically low-degree polynomials, such as linear or quadratic functions. The breakpoints or dividing points, also called knots, can be determined based on the distribution of the independent variable or prior knowledge about the relationship between the variables (Howe et al., 2016). Regression splines can be quadratic, cubic, etc. The cubic ones are especially known since normally

humans are not capable of noticing the discontinuity at each knot, which is good for the model fit. (James et al., 2013). The concept of cubic splines was actually introduced by Hastie and Tibshirani in 1990. As an example, let's see how a cubic regression spline with K knots looks like:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \ldots + \beta_{K+3} b_{K+3}(x_i) + \varepsilon_i \qquad \textbf{(3.7.)}$$

In which b_1, b_2, …, b_{K+3} are the basic functions that need to be chosen appropriately. Those use a third-degree polynomial to model each segment and ensure smoothness at the breakpoints. They do so by imposing certain continuity and smoothness constraints on the spline functions, which help to avoid drastic changes and keep a more natural representation of the underlying relationship (Hastie & Tibshirani, 1990).

A very good way of representing this is with a truncated power basis function. This equation is a good representation of one, from James et al., 2013:

$$h(x,\xi) = (x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise,} \end{cases}$$

where ξ is the knot value. When aiming to apply a cubic spline to a specific dataset with K knots, the process involves conducting a least squares regression with an intercept and 3+K predictors. These predictors take the forms X, X^2, X^3, $h(X, \xi_1)$, $h(X, \xi_2)$, … , $h(X, \xi_K)$ denote the knots (Eubank, R., 2010). Since this entails estimating K+4 regression coefficients, it can be stated that fitting a cubic spline with K knots utilizes K+4 degrees of freedom. For instance, if a term such as $\beta_4 h(x, \xi)$ is added to the model, it will introduce a discontinuity solely in the third derivative at ξ, while the function remains continuous with continuous first and second derivatives at each of the knots (James et al., 2013).

Unlike traditional regression models that assume a linear relationship, regression splines do not require a specific functional form. They offer more flexibility than other nonlinear regression techniques, such as polynomial regression, as they can fit curves with different slopes and curvatures in different segments of the independent variable's range. This flexibility allows them to handle outliers and extreme values better, as the polynomial functions are only fitted to small segments of the data. (Lin et al., 2004).

However, interpreting regression splines can be more complex than linear regression models. The relationship between the variables is described by multiple polynomial functions instead of a single equation (De Boor, 2001). Additionally, the choice of breakpoints can significantly impact the analysis results. Different choices of breakpoints, or knots, can lead to different models and conclusions. The choice of the optimal number of knots and their placement needs guarantee an equilibrium between model complexity and goodness of fit, because we want to avoid overfitting (Eubank, R., 2010).

Regression splines have been successfully applied in various fields, including economics, ecology, and epidemiology. In economics, they have been used to model the relationship between income and consumption and to estimate the effects of tax policies on consumption (Hastie et al., 2009). In ecology, regression splines have been used to model the relationship between species richness and environmental variables like temperature and precipitation (Wood, 2006). In epidemiology, they have been used to investigate the relationship between exposure to a risk factor and the occurrence of a disease. (Ruppert et al., 2003).

In conclusion, regression splines are a valuable statistical technique since they capture complex patterns in the data without requiring strict assumptions regarding a specific functional form. Besides, they are able to handle outliers and extreme values. However, their use requires careful consideration of breakpoints, and a more complex interpretation of the results compared to linear regression models. Researchers from various fields can benefit from their use when analyzing complex data Another approach that can bring numerous advantages when it comes to data modelling is GAMLSS, which we explore next.

3.3. Generalized Additive Models for Location, Scale and Shape (GAMLSS)

Moving on to Generalized Additive Models (GAM's), these constitute a more flexible statistical framework that is useful for modeling relationships between data allowing non-linear functions for the variables, while maintaining additivity (Hastie et al., 2009). GAMLSS, specifically, is a statistical modelling method that generalizes the generalised linear models (GLM's), allowing more flexible modelling of the distribution of the response variable. GAMLSS will therefore provide a better fit compared to traditional GLM's because they accommodate and capture heterogeneity in the parameters of the distribution and complex relationships between predictors and response variables (Rigby & Stasinopoulos 2005).

To do so, the key idea behind GAMLSS is to use non-parametric smoothing functions, such as splines. The response variable is assumed to follow a distribution from the exponential family, such as the Gaussian, Gamma, or Poisson distributions. The model separates the modeling of the location parameter (mean), scale parameter (variance), and shape parameter (skewness or kurtosis) of the distribution. Each parameter is modeled as a smooth function of the predictors using additive models. (Rigby & Stasinopoulos 2005).

In this context, a Box-Cox transformation can be used to stabilize variance and be closer to normality in the data. This technique can be applied in several modelling frameworks, GAMLSS is just one of those (Rigby & Stasinopoulos, 2001). This transformation corresponds to applying a power transformation to the response variable, typically denoted by λ, that determines the type of transformation. If $\lambda = 0$, it is a logarithmic transformation; if $\lambda = 1$, a simple identity transformation; for other λ values, it will correspond to different power transformations. The goal here is to find the most suitable λ that maximizes the normality or stabilization of variance in the data, which normally is the one that results in the most symmetric and approximately normally distributed residuals (Stasinopoulos et al., 2017).

This type of model provides a rich set of tools for model selection, diagnostics, and inference. It allows for the estimation of multiple distributional parameters simultaneously, making it suitable for a wide range of data types and research questions and for both quantitative and qualitative registers (James et al., 2013). Besides, these transformations, by transforming the response variable, allow for a more accurate estimation of the model parameters, leading to more reliable predictions.

4. THE PROJECT – APPLICATION

4.1. Cohort Studies

In epidemiology, a group of individuals who have a certain life experience or characteristic in common, for example, exposure to the same event that will have occurred at a certain point in time, is called a cohort (Rothman & Greenland, 1998). Speaking in epidemiological terms, it is defined as a group of people who are followed over a certain period of time, in order to assess the evolution in the development of certain outcomes, such as a disease, a variable or a condition of interest. Thus it allows to investigate the association between exposures and outcomes (Grimes & Schulz, 2002).

A cohort study can help identify risks or protective factors, thus conducting to more effective interventions and to evidence-based medicine and better public health policies. Cohort studies can be observational, in which the participant is investigated individually, through observations taken on a variable or subject at a particular point in time, or experimental, which are always prospective studies (Twisk, 2013). Prospective, or longitudinal, means that whatever the variable of interest is, it is measured repeatedly during a certain period of time. The longitudinal nature of the data allows for the establishment of temporal relationships (Setia, 2016).

In the context of this internship, we mainly inspected cohort effects, which refer to differences between groups of individuals born at different times, age effects, which refer to changes in the incidence or prevalence rates of a disease or health condition as people get older. In addition to these, there are also period effects, which refer to the change in the incidence or prevalence rates of a disease or health condition that occur at a given period of time, but this effect will not be given much relevance in this study (Twisk, 2013).

The intention beyond these studies is to understand how individuals behave in the presence of this common element in different moments, allowing to examine not only the evolution of an output variable individually but also how that variable can somehow be related to the development of other variables (Twisk, 2013). The main difference that can be drawn between the two types of studies is that in the observational ones, one can not identify what is the cause leading to certain results obtained whilst in experimental ones that issue can be answered (Twisk, 2013).

Conducting a study of this nature evolves guaranteeing that several steps are completed such as the exposure assessment with data collection phases, etc, then the follow-up, tracking the occurrence of the

results being analysed over time, the data analysis, etc (Grimes & Schulz, 2002). These types of study methods are relatively complex, as they have some challenges associated. High funding, time consumption and the necessary specialized human capital are just some of them (Porta, 2014). However, these can be particularly interesting if a country wants to better understand how its population develops and what changes are observed over time. Moreover, the information obtained through cohort studies allows many scientific questions to be answered and can provide super useful insights into health promotion, health protection and disease prevention measures (ISPUP, 2021).

4.2. The cohorts – Epiteen and Generation XXI

In this subsection, I will provide details about the project to which I devoted most of my internship time, which focused on the assessment of age and cohort effects on anthropometric measures of children/adolescents from two cohorts in the metropolitan area of Porto – EPITeen and Generation XXI.

To put this analysis into practice, I used several methods, namely the statistical techniques that were developed earlier in this report. Those were used as part of this project and were applied to the data from children in the EPITeen and Generation XXI cohorts, using the R software. Data collection in these cohorts was super relevant since it provided essential information for the adequate planning of health promotion measures and also helped achieve several scientific objectives through articles published in national and international scientific journals. A total of 2853 and 7184 observations were collected and analysed for the Epiteen and the GXXI cohorts, respectively. Details about these two cohorts and are provided below.

Epiteen – Epidemiological Health Investigation of Teenagers in Porto

EPITeen is a population-based cohort of 2788 individuals born in 1990, that was established in 2003 in Porto, Portugal. The people to be followed in this cohort attended public and private schools in the referred city. The main objective was to understand to what extent and in what way certain habits acquired during adolescence will impact on adult health. To this end, they followed the participants throughout their lives, assessing them at 13, 17, 21, 24 and 27 years, and continuing to follow them throughout life (ISPUP, 2021).

Of the 2788 students born in 1990 and enrolled in Porto's schools, 2161 adolescents participated (77.5%). The remain (627) could not be reached because they were either missing classes or did not return the information consent. Figure 4.1. explains that in more detail.

Following the aforementioned idea of a cohort study, the health professionals wanted to assess the evolution over time of some objective measures such as blood pressure, bone mineral density, weight, height and respiratory function, etc. Follow-up information is updated and completed at each assessment by taking biological samples at in presence evaluations and, in addition, collection is made through questionnaires – at home and at school - indicative of the participant's social and familiar characteristics, physical activity, alcohol intake, medical history, etc (Ramos, 2006). All the anthropometric measurements were obtained at school, by trained professionals following international guidelines.

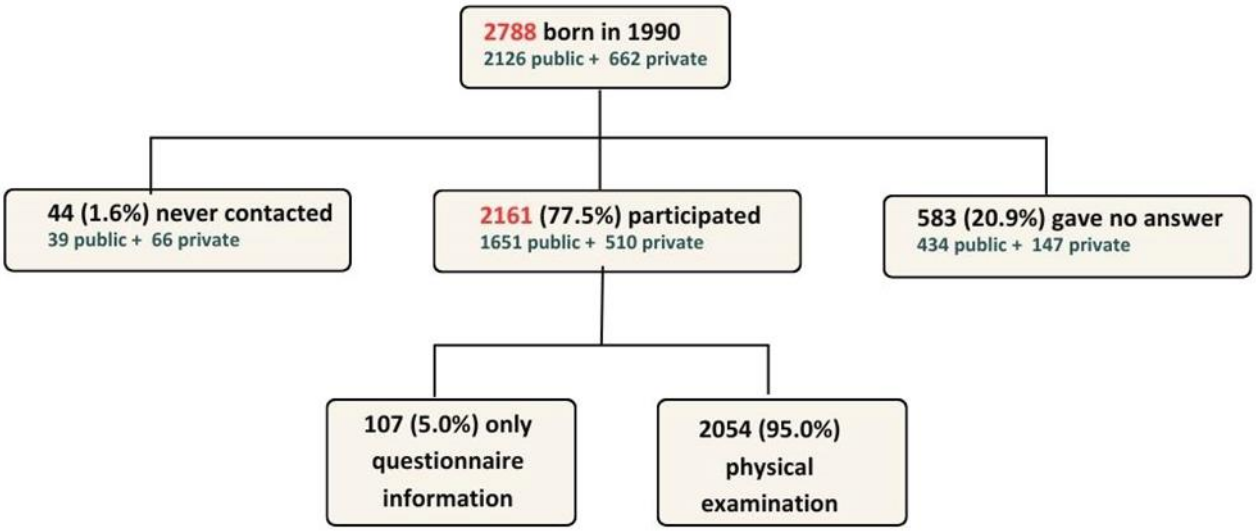


Figure 5: Baseline participation in the Epiteen cohort (Source: own elaboration)

In order to have a minimum number of students per school, all schools that had less than 4 students of each sex were excluded. At 13 years-old, from 2161 participants, 42 were not considered because they were in schools with less than 4 girls and/or 4 boys, making a total of 2119 adolescents; For the 17-year-old sample, 2257 adolescents were evaluated, but 242 were eliminated for the same reason.

Generation XXI

Generation XXI (GXXI) constitutes the first prospective population-based Portuguese birth cohort, assembled in Porto. It is focused on assessing which characteristics of pregnancy and childhood

are related to health status in subsequent phases of adolescence and adulthood, namely in which refers to the development of obesity and body composition (ISPUP, 2021). Accompanying 8647 newborns, the GXXI birth cohort was born in 2005 in the Department of Clinical Epidemiology, Predictive Medicine and Public Health of the University of Porto (Geração 21, 2020).

Children from public maternity hospitals in the city of Porto were recruited between April 2005 and August 2006, with mothers either 1) living in one of the six municipalities of the metropolitan area of Porto: Gondomar, Maia, Matosinhos, Porto, Valongo or Vila Nova de Gaia; 2) delivering at the public hospitals in those municipalities; or 3) giving birth to live babies with gestational age greater than 24 weeks.

Of the invited mothers, 91% accepted to participate, totalizing 8647 children and 8495 mothers enrolled at baseline. Children were assessed at birth, at 6, 15 and 24 months of age (n=1555, n=1043 and n=855, respectively) and subsequently at 4, 7, 10 and 13 years of age (with a participation proportion of 86%, 80%, 76% and 54%, respectively) (Costa et al., 2020). There are periods with more frequent assessments, particularly in the early years, so that information can be collected at key moments in its growth (Larsen PS et al., 2013). The evaluations were performed by face-to-face interviews or, for the families not able to participate in-person (20% and 15% at 4 and 7 years, respectively), it was performed by telephone using a shorter version of the questionnaire (Vilela, 2018). The following figure helps to visualise how it all unfolded over the years.

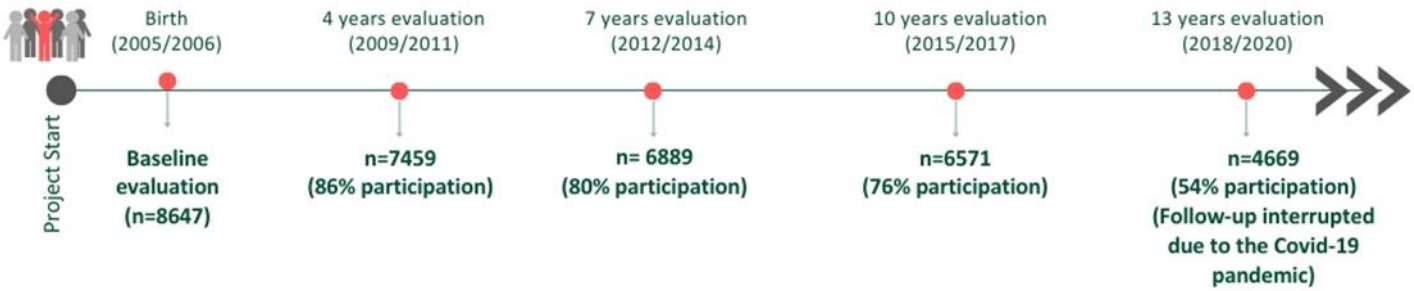


Figure 6: Participants evaluations of the birth cohort GXXI (Source: own elaboration).

Although there are naturally other European cohorts of this kind that aim to follow the development of the health status of a considerable number of children, GXXI has become particularly

relevant as it has served as the basis for a series of scientific research works in areas such as obesity, cardiovascular health, lifestyle, among others. By helping us characterise various fases of child's life and identify their determinants, these collections of information will be very useful in helping to portray the reality of health in Portugal and help to project and prepare for its future.

4.3. Data and output analysis

4.3.1. Data Cleaning

After obtaining the relevant data in the manner mentioned in the previous section, we carried out a descriptive analysis of them. I started by drawing up two tables, one for each cohort, in which I summarized the distribution of the results regarding some specific variables of interest, between participants and non-participants in the study, i.e. between individuals with known anthropometric measures and individuals without. For both cohorts, some variables of interest were chosen to be analyzed in order to give a general good approach of the data obtained. Namely the sex of the participant, the gestational age, that is, at how many weeks was born, the weight and length at birth, the number of years of schooling of the participant's mother and the information on whether or not the person was breastfed. In addition to these variables that were common to both groups of individuals, some others specific to each cohort also deserved some attention. In the case of the Epiteen, the number of years of schooling of the mother was also considered. For the GXXI, the years of schooling (number of completed years of schooling), the level of household income, in thousands, and the parity, i.e. the previous number of births, were also calculated.

Tables 1 and 2 show the results obtained concerning the p-values, which tell us whether an effect exists or not, that is, about the statistical significance of the variables. However, besides that, it was also part of this internship task to calculate the effect-size, which reveals the size of the effect, something that the p-value does not do. This effect size usually pays to be calculated for larger samples and, in this case, having been calculated only for the GXXI cohort, I chose not to include it in the tables, in order to standardize the results presented. However, this was part of the tasks performed. The tables follow:

EPITEEN	NON-PARTICIPANTS	PARTICIPANTS	P-VALUE
Sex at birth, n(%)			0.05
Female	1092 (50.1)	415 (54.3)	
Male	1086 (49.9)	349 (45.7)	
Gestational Age, M (sd)	38.6 (2.8)	39.0 (2.2)	<0.001
Weight, M (sd)	3346 (546)	3329 (484)	0.5
Length, M (sd)	49.5 (2.7)	49.4 (2.3)	0.7
Mother's schooling time, M (sd)	9.1 (4.6)	10.5 (4.6)	<0.001
Schooling time (per category), n(%)			<0.001
<=6	822 (40.3)	208 (27.4)	
7-9	368 (18.0)	138 (18.2)	
10-12	452 (22.1)	206 (27.1)	
>12	399 (19.5)	207 (27.3)	
Maximum schooling	10.1 (4.6)	11.5 (4.5)	<0.001
Breastfed, n(%)			0.1
Yes	227 (15.6)	72 (12.7)	
No	1226 (84.4)	493 (87.3)	

Table 1: Distribution of results of some variables between participants and non-participants in the Epiteen cohort (Source: own elaboration).

GXXI	NON-PARTICIPANTS	PARTICIPANTS	P-VALUE
Sex at birth, n(%)			
Female	947 (49.4)	3290 (48.9)	0.67
Male	969 (50.6)	3441 (51.1)	
Gestational Age, M (sd)	38.4 (2.1)	38.5 (1.9)	
Weight, M (sd)	3132 (562)	3153 (531)	0.07
Length, M (sd)	48.5 (2.7)	48.6 (2.5)	0.15
Mother's age, M (sd)	27.3 (6.1)	29.5 (5.4)	0.14
Schooling time, M (sd)	9.2 (3.9)	10.8 (4.3)	<0.001
Schooling time (per category), n(%)			<0.001
<=6	588 (31.0)	1450 (21.7)	<0.001
7-9	589 (31.0)	1601 (23.9)	
10-12	449 (23.7)	1851 (27.7)	
>12	271 (14.3)	1792 (26.8)	
Breastfed, n(%)			
Yes	306 (19.1)	1133 (19.3)	0.83
No	1296 (80.9)	4727 (80.7)	
Income, n(%)			
1	197 (12.6)	359 (6.1)	<0.001
2	603 (38.6)	1865 (31.8)	
3	382 (24.4)	1723 (29.3)	
4-5	284 (18.2)	1435 (24.4)	
6-7	97 (6.2)	492 (8.4)	
Parity			<0.001
0	578 (66.7)	2187 (80.6)	
1	221 (25.5)	434 (16.0)	
>2	68 (7.8)	91 (3.4)	

Table 1: Distribution of results of some variables between participants and non-participants in the GXXI cohort (Source: own elaboration).

An initial analysis of the values obtained for these variables allowed us to make some general observations. Therefore the individuals with anthropometric measures and with no measures were

compared. In both groups, the mean gestational age recorded is within what would be expected, i.e. between 37 and 42 weeks. Additionally, it is also noticeable that a large majority were not breastfed compared to those who were, for both participants and non-participants. In the Epiteen cohort, the gestational age was higher for the participants, as well as the mother's schooling time (and the maximum schooling). In the case of the GXXI, we had additional access to household income and the analysis of the results showed that the majority had average values of around 2000€ per month, for both participants and non-participants. Besides, the gestational age in this cohort was also higher for participants, as well as the mother's age, the schooling time and the measures at birth – weight and length.

Further on in the data analysis, we detected that participants' weights included outliers, i.e., implausible values. These observations being outside the expected reference values, were necessary to be eliminated. As can be seen in figure 7, on the left, through the blue points representing observations, this was essentially observed for GXXI, where outliers appear either at age 0, with weight values appearing between 50 and 100kg, or on the contrary, for ages between 6 and 12, with values below 25kg. Naturally, I decided to remove these outliers and to do so, the criterion of average weight +/- 3 standard deviations (sd) was used. This means that we considered as outliers all the values that were lower than the average weight less 3sd or higher than the average weight plus 3sd. This way, the implausible values were eliminated and we obtained the results of figure 7, on the right. were obtained, without outliers. The difference is visible essentially for GXXI.

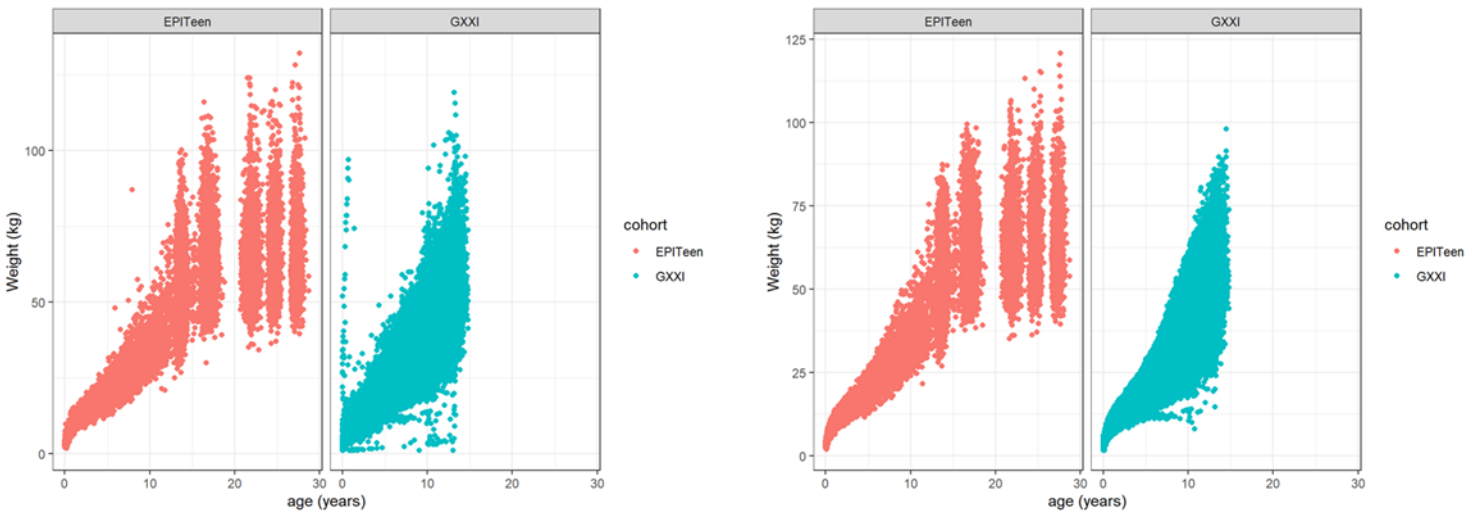


Figure 7: Before and after eliminating missing values of weight (Source: own elaboraton).

The same type of procedure was carried out for the height data, as can be seen in the graphs below. Again, the presence of outliers is more significant in the generation XXI, where there are some very low values around 10 to 12 years, as well as high values for very low ages. The criterion used was the same as the one we had used for weight (+/- 3 sd).

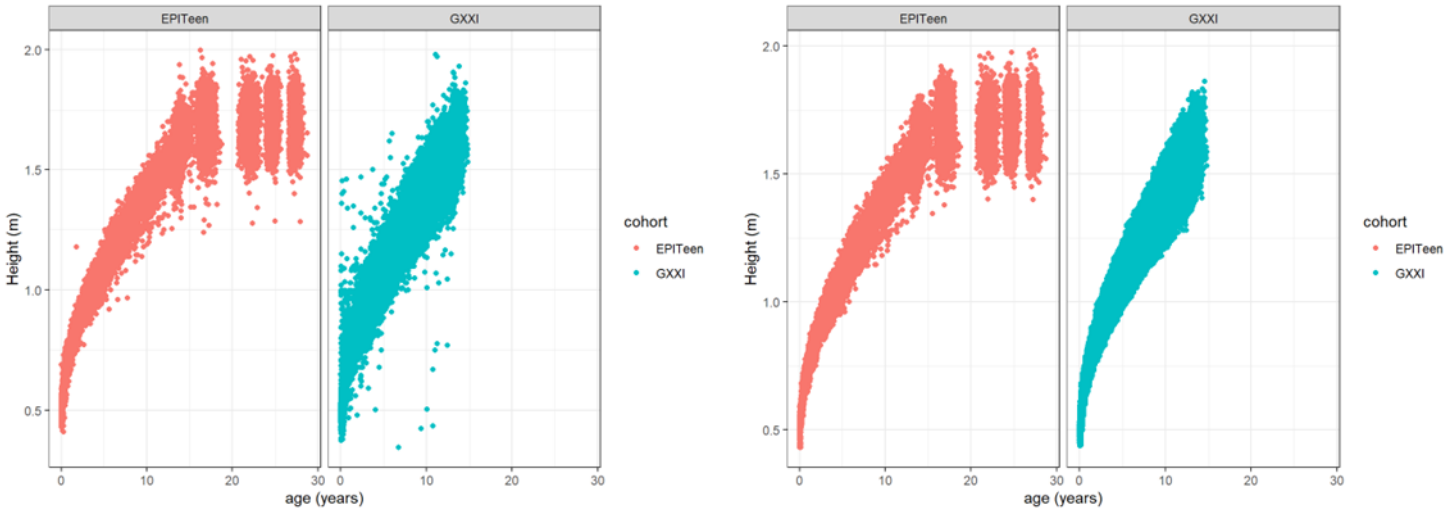


Figure 8: Before and after eliminating missing values of height (Source: own elaboraton).

Subsequently, from the values obtained for weight and height, we calculate the body mass index (bmi), which is nothing more than weight, in kg, over height, in metres squared.

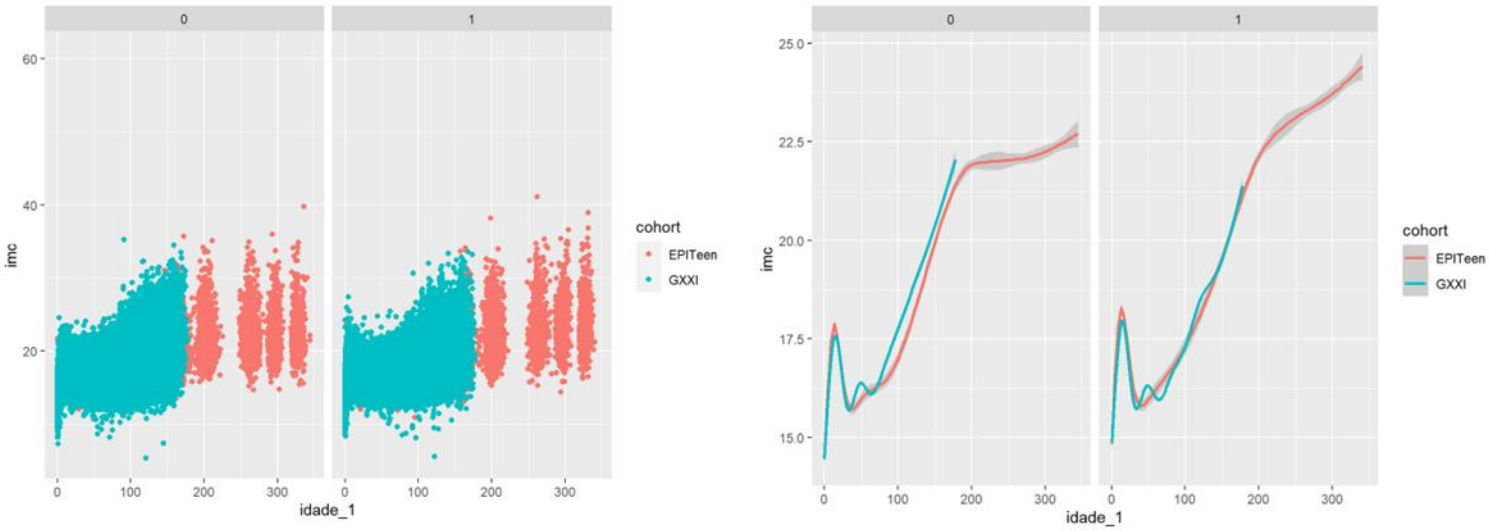


Figure 9: Before and after eliminating missing values of BMI (Source: own elaboraton).

4.3.2. Data Description

After cleaning the variables' names, eliminating missing records and removing outliers for weight, height and BMI, I proceeded to the inference analysis of the measures finally obtained, in order to estimate the relationship between cohort and age effects, which is the ultimate goal of this study. We started by calculating the mean values by year.

As it is observed in the figures below, for height and weight, there was a positive trend of these values with age, which makes perfect sense, with no significant differences between boys and girls and between the two cohorts. In the case of the BMI, there is an ascending peak in the first year of life, when growth is naturally more exponential, followed by a slight deceleration and a more linear evolution with age. Still, low peaks are recorded in the Epiteen cohort for both sexes, at around 11 years of age. It can also be observed that, for girls, the Generation XXI cohort has a constant trend slightly above the Epiteen cohort since age 6. Essentially, we could note that the average has changed over time based on the records of the two cohorts.

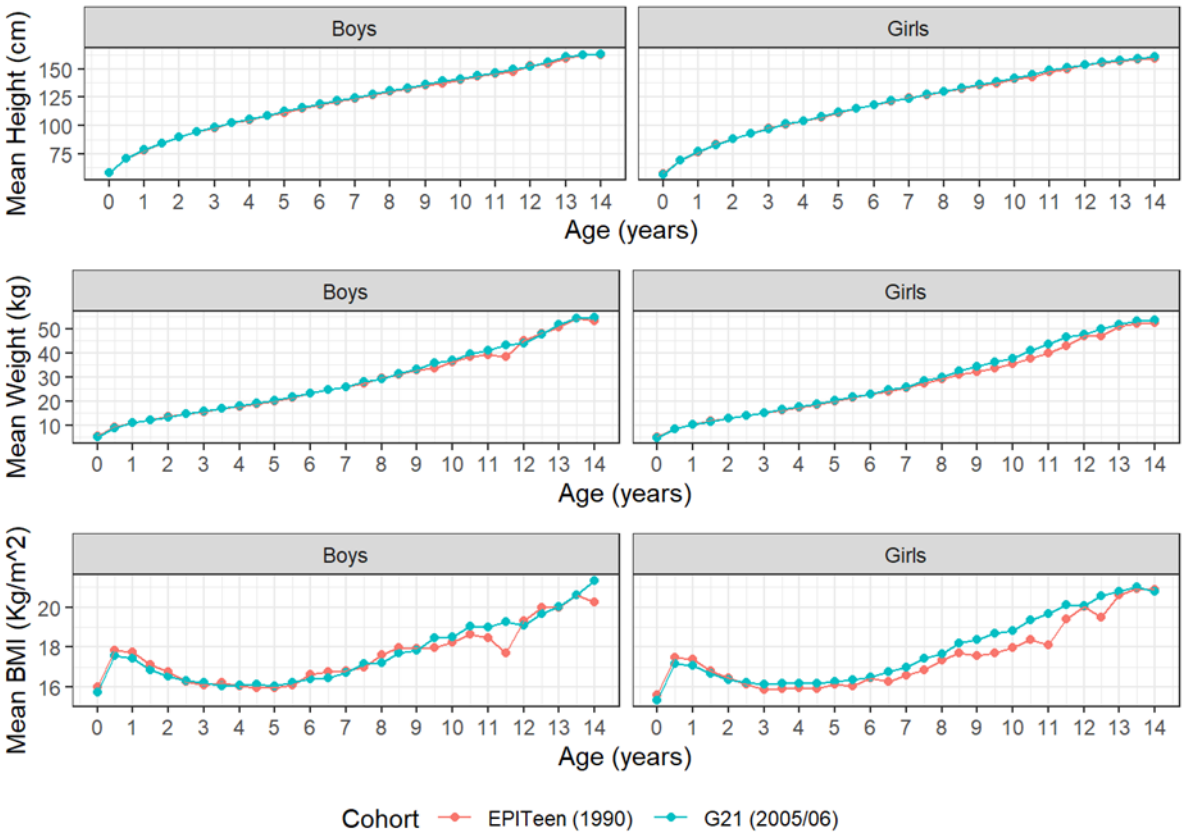


Figure 10: Data description for the mean values (Source: own elaboraton).

Next, we wanted to analyse not only the average values but also the standard deviation (SD) of the observations. To that end, we determined the coefficient of variation (CV), which is nothing more than the ratio between the standard deviation and the mean. We chose the CV because although both CV and SD measure the dispersion of values around the mean, the CV is independent of the unit in which the measurements were made, which allows me to compare the degree of variation of the data set of one cohort with that of the other, even if the two have very different mean values. We could observe that the most significant variations are registered in the first year of life, which is natural, as well as a peak, more accentuated in boys, around 12 years of age. Moreover, looking at the records, we could verify that at the most advanced ages, a greater dispersion was registered, with the GXXI cohort exhibiting greater dispersions when compared with the Epiteen one.

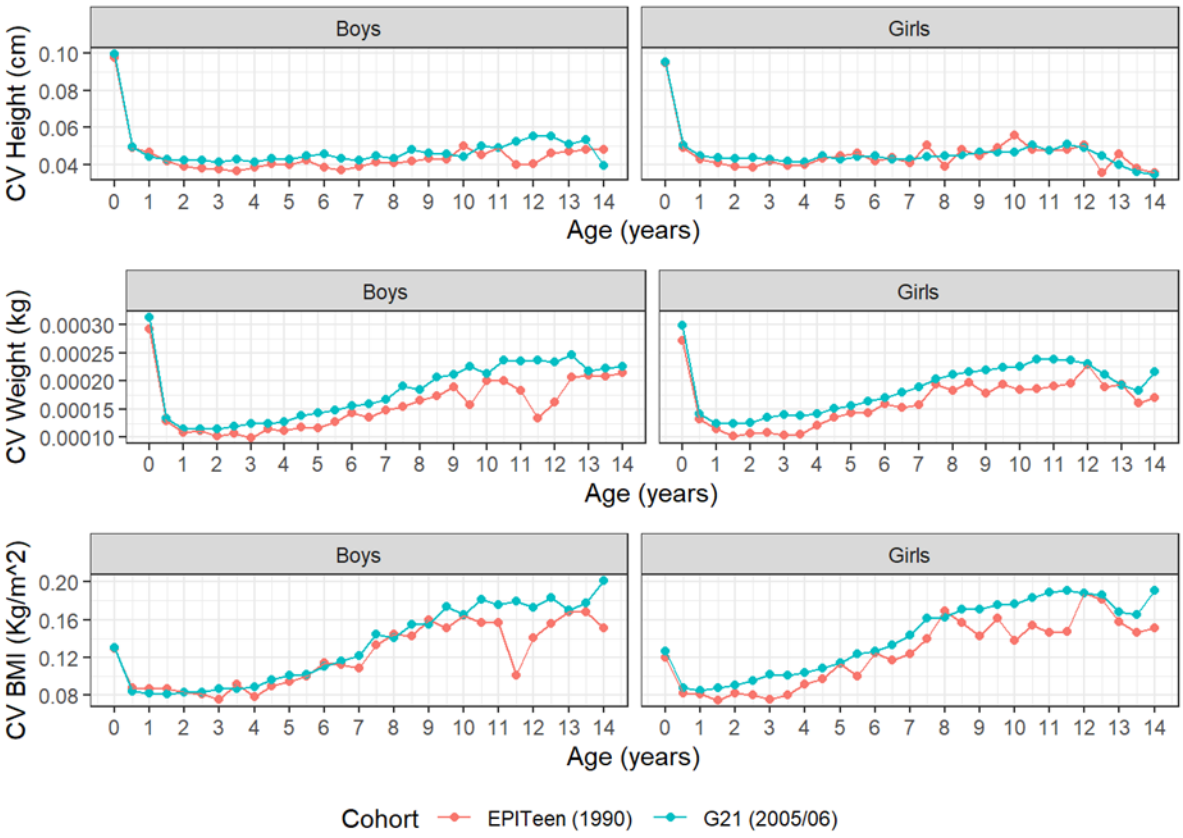


Figure 11: Data description for the CV values (Source: own elaboraton).

Following with the analysis, we proceed to the calculation of the symmetry coefficient. The data for this measurement is represented in figure 12 presented below. By determining the symmetry

coefficient for the three outcomes by year, we found that the data were not symmetrical over the age range, with records above or below zero for weight, height and BMI. If the symmetry coefficient is above or below zero, it means that there are heavy tail to the right or left, respectively. It can also be said that there is a tail on the right or left, respectively. Now, a graph with normal distribution, and therefore without tails, would have a symmetry coefficient equal to zero, being perfectly symmetrical.

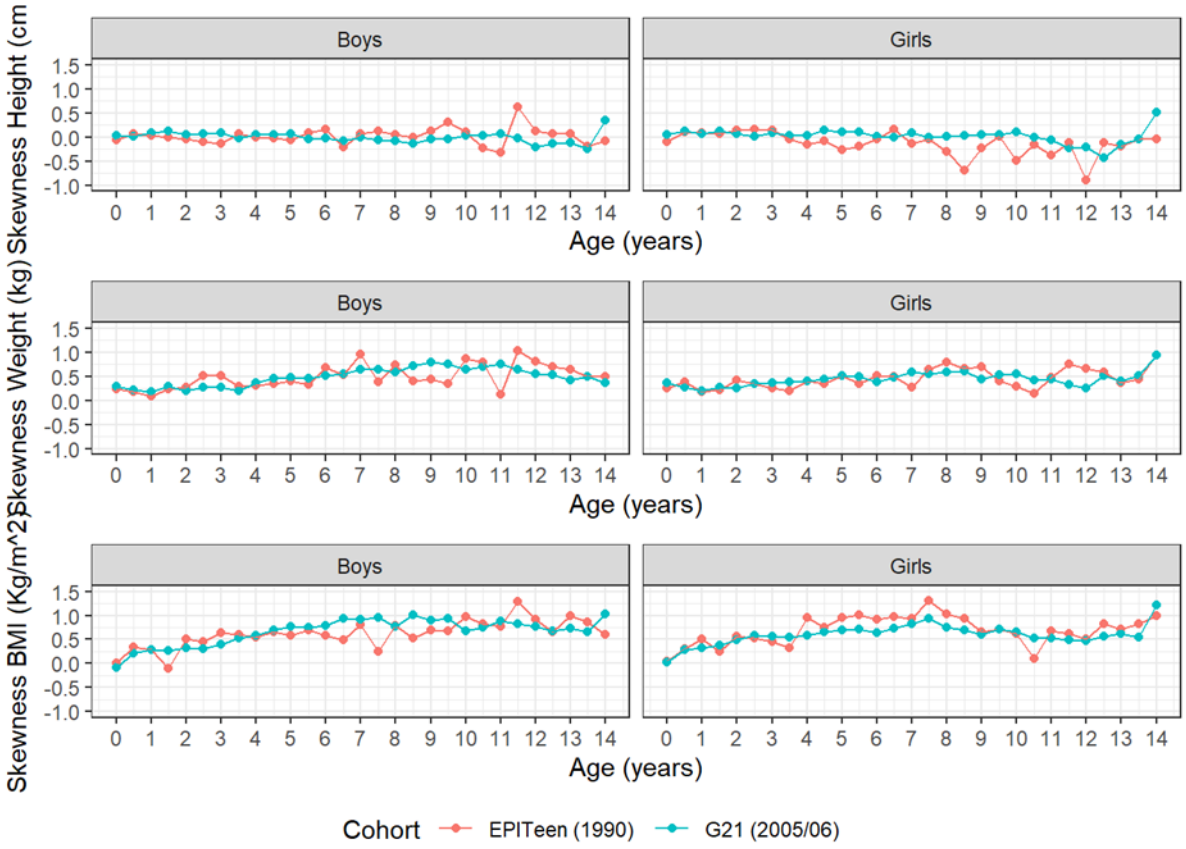


Figure 12: Data description for the skewness values (Source: own elaboraton).

In order to restore this symmetry in our data, we tried to identify which power it would be necessary to apply to them to remove this effect and be able to counterbalance the verified asymmetry. Thus, for the values of symmetry coefficient that appear above zero, for any of the variables, we would expected that the power of the symmetry coefficient of these variables would give us values below one, in order to counterbalance the initial data. In fact, this was verified as indicated in the graphs of figure 13.

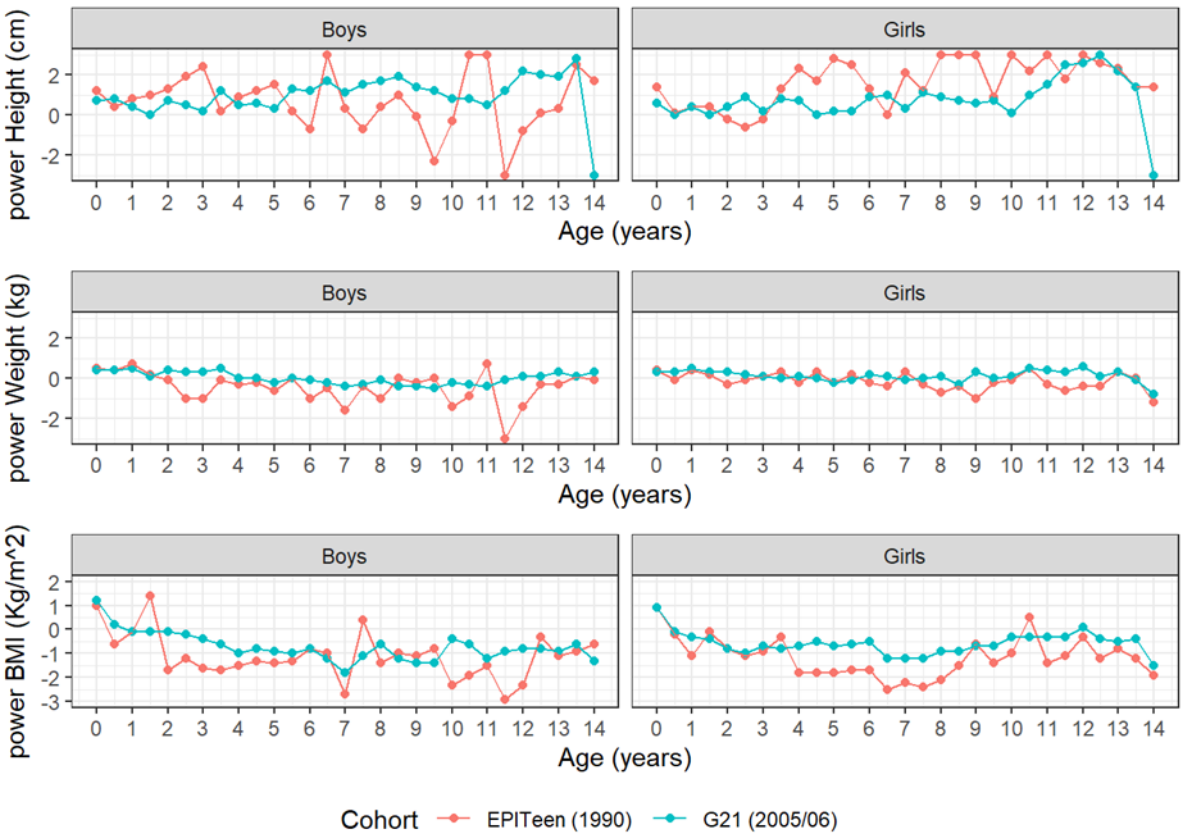


Figure 13: Data description for the power values (Source: own elaboraton).

4.3.3. Mixed-effects models with splines

After the cleaning and the descriptive analysis of the data, that allowed to have a good overview of the data sample, we reached the stage of finding the model that would make the best possible fit with them. For that purpose, to the data obtained from the cohorts, the linear mixed effects models were applied and later those were integrated with splines.

Initially, we applied the most basic model of all, in which we placed only age, that is, we assumed only age as a random variable. We call this model Model 1 and it is shown in figure 14. As expected, the fitting obtained with this Model 1 was relatively poor. In Annex 4 it can be seen in more detail the graphs obtained concerning the behaviour and the strange configuration of the observations, which was supposed to look like a cloud, but in fact did not.

PREDICTORS	ESTIMATES	CI	P-VALUE
(Intercept)	15.15	14.99-15.30	<0.001
Age effects	0.27	0.26 – 0.28	<0.001
Cohort effect [GXXI]	-0.11	-0.27 – 0.06	0.206
Age 12 months[1]	-0.36	-0.38 – -0.35	<0.001
Age 44 months[1]	0.13	0.13 – 0.14	<0.001
*Age effect * Cohort effect [GXXI]*	-0.02	-0.03 – 0.00	0.026
*Cohort effect [GXXI] * Age 12 months*	0.03	0.01 – 0.04	0.001
*Cohort effect [GXXI] * Age 44 months*	-0.02	-0.02 – -0.01	<0.001
Random Effects			
s^2	2.31		
τ_{00}	1.71		
ICC	0.43		
N	4263		
Observations	89365		
Marginal R2 / Conditional R2	0.256 / 0.573		

Table 3: Results obtained for Model 1 – Linear Case (Source: own elaboration).

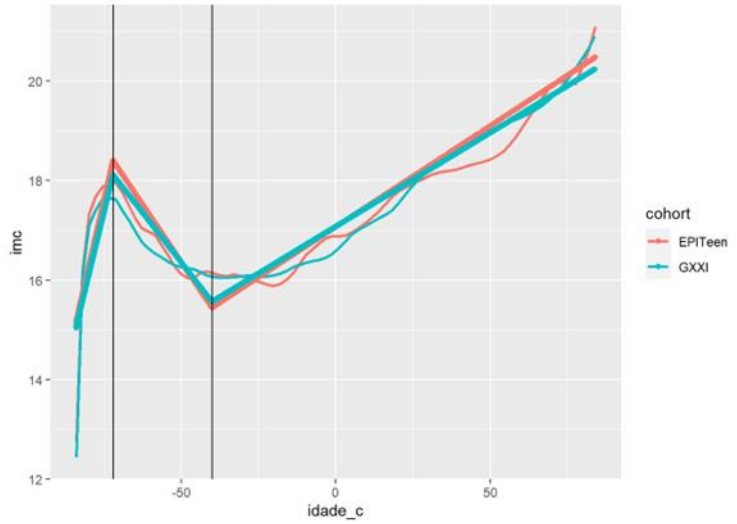


Figure 14: Model fit Model 1 – Linear Case (Source: own elaboration).

[1]The variables *Age 12 months* and *Age 44 months* mean that age was considered zero up to 12 and 44 months, respectively, starting to count the time from there.

Next, in order to improve the fit obtained so far, we introduce higher-order terms, by raising an age to a power. In this case, we add the age to the square (left side of figure 15) and also to the cube (right side of figure 15). We called it Model 2 and the representations can be found below. By doing so, a better fit would be expected since it captures eventual bends and curvatures not captured with the linear case, that we obtained in Model 1. By doing this, the fit improved but still had some imperfections. Tables 4 and 5, show the model details obtained raising it to the square, that is the model with the quadratic age, and the cubic age, respectively. We also included the splines, for which it was necessary to choose the most appropriate knots, that is, the breakpoints that best fitted the data distribution. In a trial-error logic and after testing several possibilities for the knots values, we concluded that those that resulted in a better model, were 12 and 44 (months). Those are are precisely the moments in the graphs of figure 15 in which the inflection points can be noted.

PREDICTORS	ESTIMATES	CI	P-VALUE
(Intercept)	-205.95	-214.66 – -197.24	<0.001
Age	-6.08	-6.32 – -5.84	<0.001
Cohort effect [GXXI]	7.04	-2.02 – 16.09	0.128
Age²	-0.04	-0.04 – -0.04	<0.001
Age 12 months²	0.04	0.04 – 0.05	<0.001
Age 44 months²	-0.00	-0.00 – 0.00	<0.001
*Age effect * Cohort effect [GXXI]*	0.19	-0.06 – 0.44	0.130
*Cohort effect [GXXI] * Age²*	0.00	-0.00 – 0.00	0.137
*Cohort effect [GXXI] * Age 12 months²*	-0.00	-0.00 – 0.00	0.134
*Cohort effect [GXXI] * Age 44 months²*	0.00	-0.00 – 0.00	0.311
Random Effects			
s^2	2.03		
τ_{00}	1.75		
ICC	0.46		
N	4263		
Observations	89365		
Marginal R2 / Conditional R2	0.303 / 0.626		

Table 4: Results obtained for Model 2 – Quadratic Case (Source: own elaboration).

PREDICTORS	ESTIMATES	CI	P-VALUE
(Intercept)	1331.19	1263.96 – 1398.41	<0.001
Age	54.79	52.00 – 57.57	<0.001
Cohort effect [GXXI]	21.36	-48.61 – 91.33	0.550
Age²	0.76	0.72 – 0.80	<0.001
Age³	0.00	0.00 – 0.00	<0.001
Age 12 months³	-0.00	-0.00 – -0.00	<0.001
Age 44 months³	0.00	0.00 – 0.00	0.029
*Age effect * Cohort effect [GXXI]*	0.87	-2.03 – 3.77	0.556
*Cohort effect [GXXI] * Age²*	0.01	-0.03 – 0.05	0.564
*Cohort effect [GXXI] * Age³*	0.00	-0.00 – 0.00	0.572
*Cohort effect [GXXI] * Age 12 months³*	-0.00	-0.00 – 0.00	0.546
*Cohort effect [GXXI] * Age 44 months³*	0.00	-0.00 – 0.00	0.154
Random Effects			
s^2	1.96		
τ_{00}	1.77		
ICC	0.48		
N	4263		
Observations	89365		
Marginal R2 / Conditional R2	0.317 / 0.641		

Table 5: Results obtained for Model 2 – Cubic Case (Source: own elaboration).

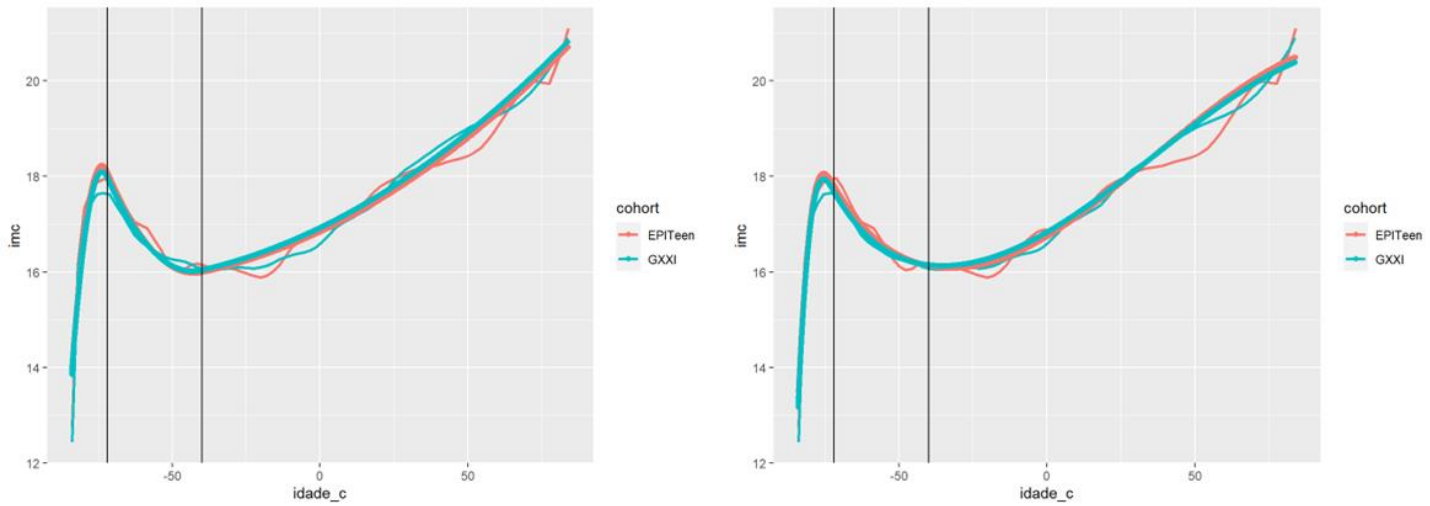


Figure 15: Model fit in model 2 - Quadratic case (on the left) and cubic case (on the right) (Source: own elaboration)

4.3.4. GAMLSS

Turning now to the concrete application of the GAMLSS model, in this context, I used a Box-Cole Green which performs a Box-Cox transformation as part of the modelling process. This involved applying a transformation power to the response variable, in order to further improve the fit of the model so far obtained, by finding the power that results in the greatest symmetry and the most normally distributed residuals. By doing this and transforming the variable, we were able to obtain more accurate estimates of the model parameters, which will potentially lead to more reliable inferences and predictions. Thus, we performed this same procedure for girls and boys and the results of these transformations are in the tables below, which show us that there is indeed an interaction between the age effect and the cohort effect, as is observable with all significant p-values. The results for girls and boys, respectively, follow:

MEAN COEFFICIENTS	ESTIMATE		STD. ERROR		T-VALUE		PR(>|T|)	
(Intercept)	15.93	16.33	0.03	0.03	644.78	554.11	$< 2e^{-16}$	$< 2e^{-16}$
pb(age, 7)	0.01	0.01	0.00	0.00	25.62	17.41	$< 2e^{-16}$	$< 2e^{-16}$
cohort GXXI	-0.29	-0.23	0.03	0.03	-10.87	-7.54	$< 2e^{-16}$	$< 4.7e^{-14}$
*pb(age, 7)*cohort GXXI*	0.01	0.00	0.00	0.00	14.29	5.25	$< 2e^{-16}$	$< 1.53e^{-07}$

Table 6: Box-Cox transformation – Mean values for girls - left columns - and boys - right columns (Source: own elaboration).

CV COEFFICIENTS	ESTIMATE		STD. ERROR		T-VALUE		PR(>|T|)	
(Intercept)	-2.34	-2.29	0.01	0.01	-273.49	-248.56	$< 2e^{-16}$	$< 2e^{-16}$
*poly**(age, 2)*	40.84	28.42	2.18	2.36	18.75	12.06	$< 2e^{-16}$	$< 2e^{-16}$
poly(age, 2)[2]	11.13	23.67	2.10	2.24	5.29	10.56	$< 1.23e^{-07}$	$< 2e^{-16}$
cohort GXXI	0.14	0.04	0.01	0.01	15.69	4.15	$< 2e^{-16}$	$3.39e^{-05}$
poly(age, 2) cohort GXXI*	20.85	24.75	2.31	2.48	9.01	9.97	$< 2e^{-16}$	$< 2e^{-16}$
poly(age, 2)[2]** cohort GXXI*	2.78	10.28	2.26	2.39	1.23	4.30	0.218	$1.70e^{-05}$

Table 7: Box-Cox transformation – CV values for girls - left columns - and boys - right columns (Source: own elaboration).

[2]The variable *poly* means that it is a polynomial quadratic term.

POWER COEFFICIENTS	ESTIMATE		STD. ERROR		T-VALUE		PR(>|T|)	
(Intercept)	-0.66	-0.36	0.09	0.09	-6.80	-3.70	$1.05e^{-11}$	0.0002
poly(age, 2)	-202.49	-222.25	22.09	21.64	-9.17	-10.27	$< 2e^{-16}$	$< 2e^{-16}$
poly(age, 2)²	148.79	133.51	19.79	20.17	7.52	6.62	$5.64e^{-14}$	$3.6e^{-11}$
cohort GXXI	0.57	0.37	0.10	0.10	5.62	3.66	$1.97e^{-08}$	0.0003
poly(age, 2) cohort GXXI*	91.52	42.29	22.90	22.43	3.99	1.89	$6.45e^{-05}$	0.0594
poly(age, 2)² cohort GXXI*	-9.02	27.94	20.87	21.33	-0.43	1.31	0.666	0.1903

Table 8: Box-Cox transformation – Power values for girls - left columns - and boys - right columns (Source: own elaboration).

As can be seen in table 6, the coefficients of the mean for cohort GXXI of -0.29 tells us that cohort GXXI has 0.29 less at the intercept, i.e. at birth, which denotes that its individuals are born thinner. Moreover, we find that there is a significant interaction between age and cohort. In the case of the coefficient of variation, described in table 7, we could conclude that initially there is a higher variance at 0.14 and again there is a significant interaction between the polynomial quadratic poly term and the cohort effect. Moving on to the symmetry coefficient it was found that the GXXI cohort has a higher power at 0.56 and again there is a significant interaction between the cohort effect and power.

These results allow us to conclude that, with this model, we were able to obtain what we vizualized previously in the descriptive analysis. However, more objectively, although we are able to draw some interesting conclusions from the values obtained in the above tables, namely the ones described above, in some cases it becomes difficult to extract interpretations since certain values are not directly interpretable. For this reason, the descriptive graphs of the above results are included below, with which it is easier to visualize the behaviour of the variables, starting with the average.

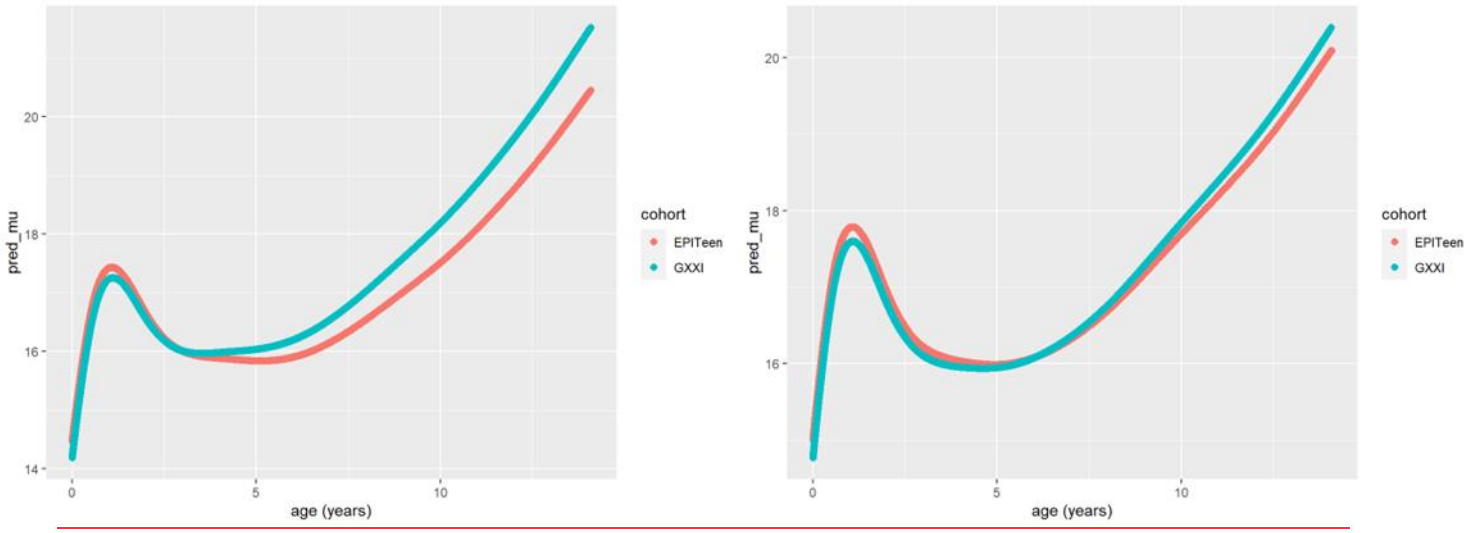


Figure 16: Description of the mean values using GAMLSS | Female (on the left) and male (on the right) (Source: own elaboration).

As we could conclude in the previous section when analysing the mean values in table 6, the individuals from GXXI with 0.29 less in the in-between - for age=0 - i.e., these children are born thinner. In this sense, in figure 17 which describes the behaviour of the mean, a more significant difference between the two cohorts should be noted, specifically at the beginning of the axes, highlighting this inferior performance of the GXXI. The reason why in figure 17 this difference is not very noticeable is that these analyses were carried out on a logarithmic scale. Naturally, at birth it is significantly different to be 3100g or 3200g. However, as we grow and the proportions of the values reached become much larger, and where the variations are much more significant, those initial variations become barely perceptible in the graph made in logarithmic scale, making it seem that the differences are small, and they're not that small. As we can also observe in figure 17, from age five onwards there is a clear separation between the evolution of age by cohort, with GXXI always being above Epiteen and therefore its participants being heavier.

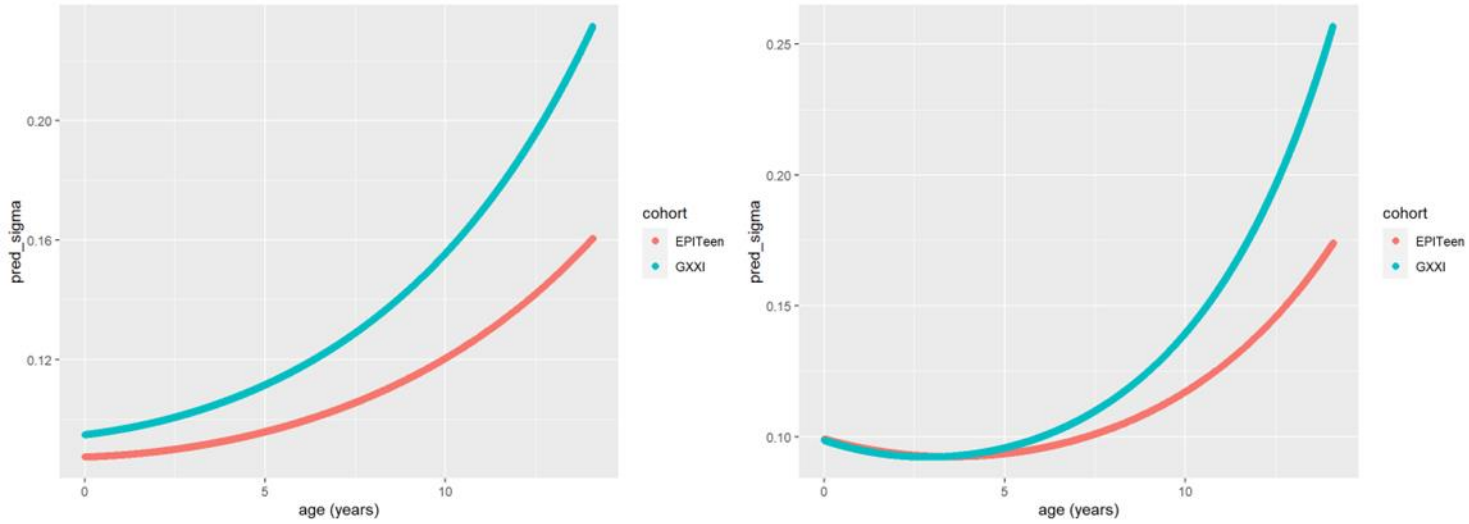


Figure 17: Description of the CV values using GAMLSS | Female (on the left) and male (on the right) (Source: own elaboration).

Moving forward to the coefficient of variation, as shown in figure 18 and according to the results in table 7, there was a higher variance in GXXI initially. It is also visible the significant interaction existing between the poly variable and the cohort effect, verifying an increasing variance with age, and keeping the variance of the GXXI higher. Specifically, in the case of boys, this difference starts to be more significant after 5 years of age.

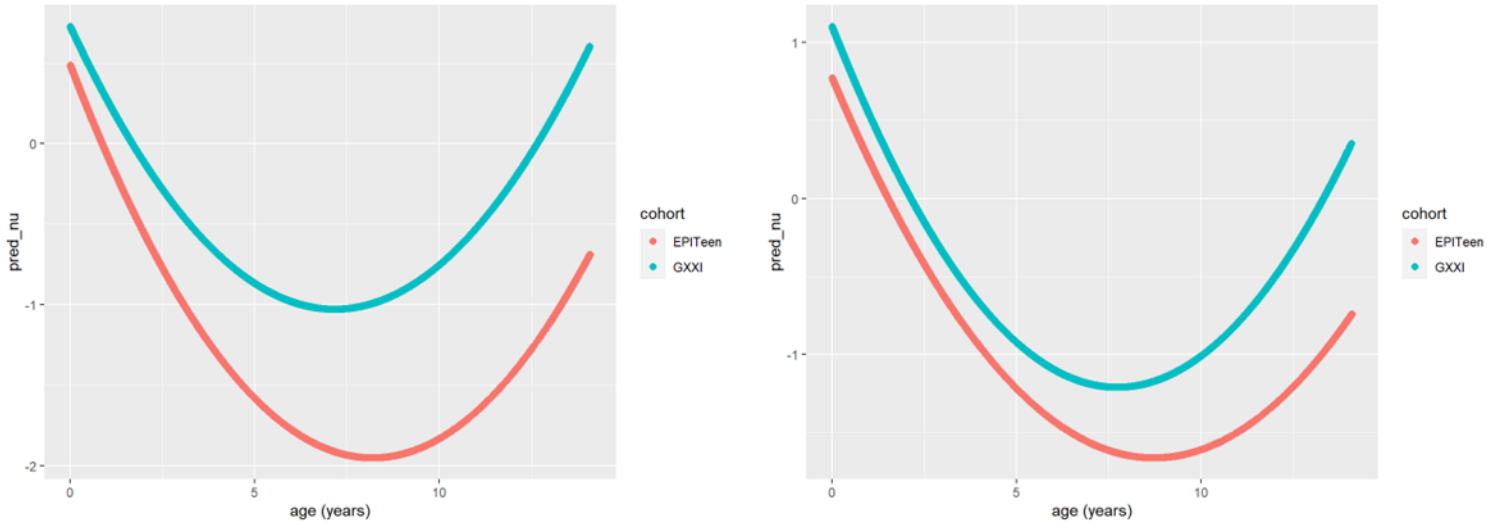


Figure 18: Description of the power values using GAMLSS | Female (on the left) and male (on the right) (Source: own elaboration).

Finally, the description of the graphics of figure 19 regarding the symmetry coefficient, shows that the cohort GXXI initially presents a higher power, according to the information analysed in table 8, for both girls and boys. Again, an interaction of the cohort effect with power was also detected.

Also, to make these results a little more interpretable, I elaborated a table on the prevalence of obesity in boys and girls, by age, through which it was possible to do a practical application of the information in the graphs presented above. What we did was to take the cut-off of the Epiteen cohort, which gave us the 95th percentile and with this cut-off, we went to estimate what the prevalence would be in the GXXI cohort, by age. Considering the cut-off point of the Epiteen cohort, which corresponded to 5%, what would be expected was that this would also be the case in the GXXI. In table 9 below, we observe that this still occurred in the first year of life, at 12 months - values of 3.1% for boys and 5.4% for girls, within the expected and in accordance with the previous results that the GXXI participants are even born thinner. From then on, these values increase more and more the difference in relation to the expected 5%. What was found was that, for the same Epiteen cut-off point, the prevalences in GXXI are all much higher, standing out especially after 108 months, or 9 years, when values of practically double that can be seen, especially in girls.

AGE (IN MONTHS)	BOYS		GIRLS	
12	20,87	3,1%	20,24	5,3%
24	19,98	4,0%	19,51	6,0%
36	19,24	4,9%	18,86	6,8%
48	19,24	5,9%	18,97	7,5%
60	19,39	6,7%	19,28	8,1%
72	19,75	7,5%	19,69	8,7%
84	20,29	8,2%	20,27	9,2%
96	20,93	8,7%	20,99	9,8%
108	21,59	9,0%	21,79	10,4%
120	22,23	9,3%	22,62	11,3%
132	22,89	9,4%	23,54	12,4%
144	23,52	9,5%	24,54	13,7%
156	24,15	9,4%	25,53	15,4%
168	24,72	9,4%	26,47	17,4%

Table 9: Analysis of obesity prevalence in both sexes.

5. CONCLUSIONS

5.1. Reflection and limitations

This internship project carried out at ISPUP was very enriching since it allowed me to explore areas outside what I was used to, and it was a rewarding challenge on several levels. As this internship was simultaneously a curricular internship in the scope of the Master in Economics of Business and Strategy and an internship integrated in the scope of the international network QTEM, it should have a highly quantitative nature where I could put into practice the acquired skills namely around the themes of Quantitative Methods, Big Data Analytics, Decision Support and Management, etc, which it definitely had.

In fact, by integrating the biostatistics team of ISPUP, more specifically in the EPIUnit, I had the opportunity to be given work with a quantitative and statistical focus, which corresponded exactly to the main objective of my internship and my duties. Therefore I can conclude that all my academic goals were fulfilled with the completion of the tasks described in this report.

At the time of writing this report, ISPUP manages the organisation of ten research cohorts. This specific study to which this report refers to uses data from two of these cohorts, which are two of the oldest of this set - EPITeen and Generation 21 - that have been followed longitudinally since 2003 and 2005, respectively. A thorough descriptive analysis of the data provided insight into the distribution of specific variables of interest among study participants and non-participants, indicative of their financial, social and academic realities, in order to access possible bias. Next, the cleaning process involved the elimination of atypical values and missing records, ensuring the reliability of the datasets. And the subsequent inference analysis, employing robust statistical techniques, aimed to estimate the relationship between cohort and age effects, which was the ultimate goal of the study.

An extensive literature review was conducted, focusing on the major topics that would give statistical contribution to the present study, namely linear mixed-effects models, regression splines and GAMLSS. An attempt was made to establish a chronological relationship between the literature, in order to identify key features that could be related to each other. In addition, it was ensured that the publications and references used to extract information came from reliable sources whose content was reviewed and considered reliable. We tried to include publications from well-known journals, articles or books whose

authors have a vast repertoire of publications in this area and on the topics mentioned. The articles range between the years 1982 and 2022, with the majority being published from 2005 onwards, showing a growing trend for researchers to analyse this topic. This part of the research and the elaboration of the literature review were very important to consolidate concepts and ideas in depth, allowing me to be much more comfortable moving on to a practical application of the same to the available data.

At the same time, this report also describes how this knowledge could be applied in practice. The findings revealed positive trends in height and weight with age, consistent with expectations. No significant differences were observed between boys and girls or between the two cohorts in terms of these variables. In the case of BMI, an ascending peak during the first year of life followed by a more linear evolution with age was observed. The Epiteen cohort exhibited low peaks around 11 years of age, while the Generation 21 cohort displayed a slightly higher trend since age 6, particularly in girls.

Additionally, in terms of examination of dispersion of observations around the mean, the most significant variations were observed in the first year of life, with a more pronounced peak around 12 years of age, especially among boys. At advanced ages, greater dispersion was registered, with the Generation 21 cohort showing higher dispersions compared to the Epiteen cohort. Furthermore, the process to improve model fit and find the best parameterization model for the data evolved the application of linear mixed effects models initially, with which I was able to notice some clear differences between cohorts in terms of mean, followed by their integration with splines. Then using the GAMLSS, we were able to add that those differences are also significant in terms of variance and symmetry. This constitutes an additional advantage associated with the use of GAMLSS, since it allowed us to detect more differences in other parameters.

Overall, this project underlines the importance of studying the long-term impact of adolescent habits on adult health outcomes and even understanding health status in childhood, adolescence and adulthood. Indeed, the main contribution of this internship and its associated project is its practical component and its public usefulness. The use of appropriate statistical techniques was crucial to obtain meaningful knowledge. The results of this research contribute to evidence-based decision-making, informing the development of interventions and policies aimed at improving public health and well-being.

Although the objectives of the internship were achieved, some limitations may also be pointed out, with a view to future improvements in this field of research, namely regarding data collection methods. Gathering records from such a large number of individuals and for such a long period of time is a big challenge and obviously implies high costs, whether in terms of human resources, or for the concrete and precise collection of such a large amount of information, or to ensure the reliability of the data, etc. A wide range of data collection methods could have been used to pool data from the different cohorts. In this case, questionnaires and face-to-face anthropometric measurements were the ones used but I believe there is a need to develop methods that are more advanced and adaptable to the increasingly automated and virtual reality. There is a need to have procedures suitable for pooled analyses in epidemiology, and also for considering which eventually different outcomes should be collected for similar cohort projects in the future.

In addition, there are some challenges that I think may create difficulties in trying to harmonise measures across both cohorts. Dealing with more than one database, with different participants, at more than one point in time and who are embedded in different contexts and projects, it could be the case that one cohort got funding to use a more up-to-date and therefore more expensive tool to obtain a measure, while the other only had funding for a much cheaper one, which could be unfair in some ways.

Also, it is important to note that within the scope of the internship, in parallel with what is presented in this report, a GAMLSS model including random effects was also developed. But what happened is that, in practice, running such a model in R became unaffordable since the processing time in order to obtain the outputs was very long, due to the relatively large size of the sample. Thus, another limitation of this study could be the fact that random effects were not included here, as doing so would make the whole process much slower.

It would be interesting if more studies were done on cohort and age effects, and the link between the two, as it is really interesting to understand how this type of information evolves with generations and throughout each generation. I believe these to be extremely useful topics, even from the perspective of creating a society that is more informed and prepared for what will come with future generations. Although studies of this kind have already been developed in several cities in Europe, I think it would be a very relevant theme to be more studied at a global level, allowing comparisons with greater significance and generality.

Indeed, a cohort effect was detected and it reflected in an increase in the prevalence of obesity. The increased prevalence of obesity in the GXXI is primarily driven by sedentary lifestyle issues/lack of physical activity and unbalanced diets. It is known that eating habits and physical activity are essential in the prevention of overweight and obesity (An R., 2017). In fact, analyzing physical activity practices, at ten years of age, only 14.1% of the participants in the GXXI cohort declared practicing physical activity (Costa et al., 2020), a really low number (and in accordance to the results obtained in Table 9). This will have a high burden in terms of the disability-adjusted life year (DALY), which is a measure of how much a disease affects the lives of a population. This will be reflected in high costs to the health economy in order to prevent it.

This epidemiological study which I gave my contribution to adds to a number of existing knowledge at the level of public health research in Portugal. In general, this type of study is an essential tool for planning health promotion, health protection and disease prevention measures.

BIBLIOGRAPHIC REFERENCES

An, Ruopeng. (2017). Diet quality and physical activity in relation to childhood obesity. *International journal of adolescent medicine and health*, *29*(2). DOI: 10.1515/ijamh-2015-0045.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. DOI: 10.18637/jss.v067.i01.

Brown, H., & Prescott, R. (2015). *Applied mixed models in medicine*. John Wiley & Sons.

Costa, A., Severo, M., Vilela, S., Fildes, A., Oliveira, A. (2020). Bidirectional relationships between appetitive behaviours and body mass index in childhood: a cross-lagged analysis in the Generation XXI birth cohort. *European Journal of Nutrition*, *60,* 239-247. DOI:10.1007/s00394-020-02238-9.

Costa, D., Severo, M., Fraga, S., Barros, H., & Lopes, C. (2018*).* Are maternal weight and placental weight inter-related? A systematic review of associations. *Maternal and Child Health Journal*, *22*(2), 222-229.

Dangerfield, C., Fenichel, E. P., Finnoff, D., Hanley, N., Heap, S. H., Shogren, J. F., & Toxvaerd, F. (2022). Challenges of integrating economics into epidemiological analysis of and policy responses to emerging infectious diseases. *Epidemics*, *39*, 100585.

De Boor, C. (2001) *A Practical Guide to Splines* (Revised Edition). Springer-Verlag.

Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. (2002). *Analysis of longitudinal data.* Oxford University Press.

Eubank, R. L. (2010). *Nonparametric Regression and Spline Smoothing* (2nd Edition). CRC Press.

Geração 21. (2020). Who we are. *Geração 21*. Retrieved June 2 2023, from https://www.geracao21.com/pt/

Grimes, D. A., & Schulz, K. F. (2002). *Cohort studies: marching towards outcomes.* The Lancet, 359(9303), 341-345.

Hastie, T., Tibshirani, R. (1990). *Generalized additive models* (Vol 43). CRC Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: springer.

Howe, L. D., Tilling, K., Matijasevich, A., Petherick, E. S., Santos, A. C., Fairley, L., ... & Lawlor, D. A. (2016). Linear spline multilevel models for summarising childhood growth trajectories: a guide to their application using examples from five birth cohorts. Statistical *methods in medical research, 25*(5), 1854-1874.

Instituto de Saúde Pública da Universidade do Porto (ISPUP). 2021. Available at https://ispup.up.pt/. Accessed at April 30, 2023.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013*). An introduction to statistical learning: with Applications in R.* (Vol. 112, p. 18). New York: springer.

Laird, N.M. and Ware, J.H. (1982). *Random effects models for longitudinal data.* Biometrics, 38, 963–74

Larsen PS, Kamper-Jørgensen M, Adamson A et al. (2013*). Pregnancy and birth cohort resources in europe: a large opportunity for aetiological child health research*. Paediatr Perinat Epidemiol, 27, 393–414. DOI: 10.1111/ppe.12060

Lin, X., Wang, N., Welsh, A., & Carroll, R. (2004). *Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. Biometrika*, *91*(1), 177-193.

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering, 48*, 500-506.

Porta, M. (2014). *A dictionary of epidemiology* (6th ed). Oxford University Press.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/

Ramos, E. (2006). *Health Determinants in Porto Adolescents: the EPITeen Cohort*. PhD Thesis. Porto, Portugal: Medical School of the University of Porto.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Rigby, R.A., & Stasinopoulos, D.M. (2001). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, *1*(3), 209-229.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized *Additive Models for Location, Scale and Shape. Journal of the Royal Statistical Society Series C: Applied Statistics, 54*(3), 507–554.

Rothman, K., & Greenland, S. (1998). *Modern Epidemiology, 2nd Edition*. Lippincott Williams & Wilkins.

Rothman, K.J., Greenland, S. and Lash, T. (2008). *Modern epidemiology* (Vol. 3). Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 303-327.

Ruppert, D., Wand, M.P., & Carroll, R.J. (2003). *Semiparametric Regression* (No. 12). Cambridge University Press.

Setia, M. S. (2016). Methodology series module 1: Cohort studies. *Indian Journal of Dermatology, 61*(1), 21-25.

Stamuli, E. & Panagiotakos D. (2022). The Role of epidemiology research in economic evaluation for Health Technology Assessment. J Atherosclerosis Prev Treat, 13(3), 96–102. DOI: 10.53590/japt.02.1038

Stasinopoulos, D.M., Rigby, R.A., Heller, G.Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing: Using GAMLSS in R.* CRC Press.

Szklo, M., Nieto F.J. & Miller D. (2001). Epidemiology: Beyond the Basics. *American Journal of Epidemiology,* Vol. 153(8), 821-822. DOI: 10.1093/aje/153.8.821

Twisk, J. W. (2013). *Applied longitudinal data analysis for epidemiology: a practical guide.* Cambridge University Press.

Verbeke, G., Molenberghs, G., & Verbeke, G. (1997). *Linear mixed models for longitudinal data* (pp. 63-153). Springer New York.

Vilela, A.S.M. (2018). *Tracking the acquisition of eating habits in children and its effects on behaviours related to appetite and on adiposity.* PhD Thesis. Porto, Portugal: Medical School of the University of Porto.

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications.* arXiv preprint arXiv:1308.5499.

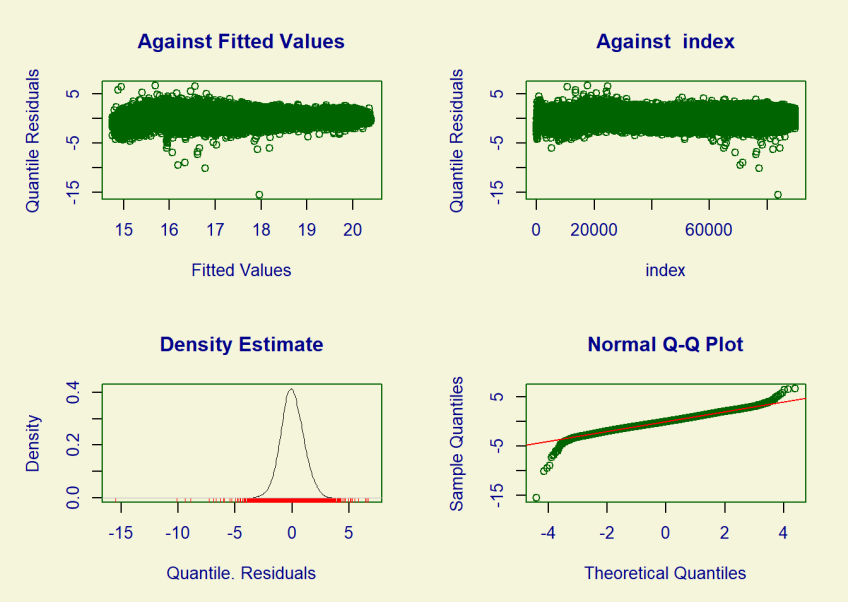
Wood, S. N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, *48*(4), 445-464.

Zhang H. (2004). Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. *Statistical Methods in Medical Research*. 13(1), 63-82. DOI: 10.1191/0962280204sm353ra

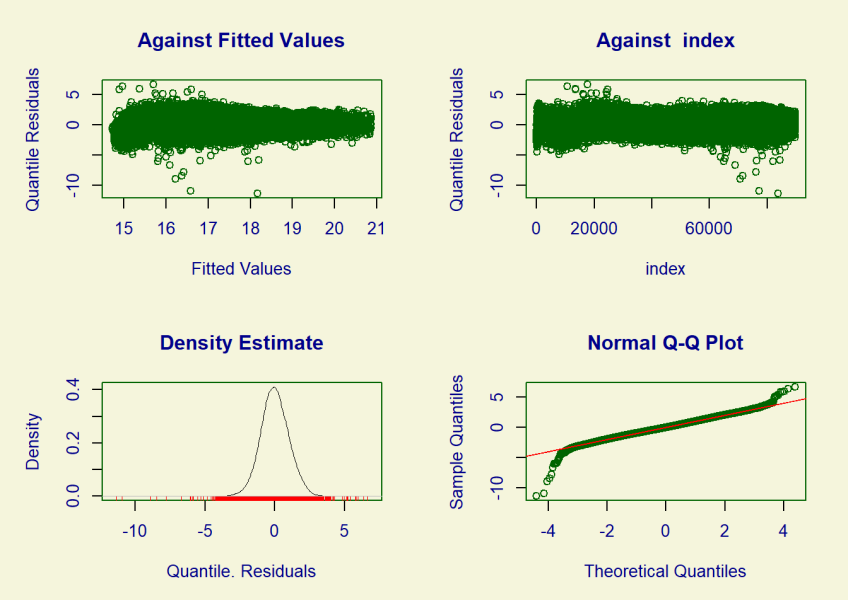
ANNEXES

February and March	- Cleaning of the data obtained from the public and private schools from the Epiteen cohort: organization by school and by group of schools in the city of Porto, in R. - Construction of tables 1 and 2 present in this report, in R, upon request of the Integrated Doctoral Researcher Joana Araújo, responsible for this research project on the effects of cohort and age in the Epiteen and GXXI cohorts. - Recurring meetings, usually weekly, with the internship supervisor and Prof. Dr. Joana Araújo, to monitor the status of the study and learn what sequence of procedures must take place in a study of this type in terms of data analysis.
April	- Focus essentially on the construction of spline models, testing the different possible knots and analysis of the resulting models. - Theoretical sessions with Prof. Milton to better understand certain important themes, namely GAMLSS and Mixed-effects models. - Meetings with the internship supervisor and Professor Joana Araújo on the 9th and 22nd of the month, similar to the ones in February and March. - Attendance at the scientific seminars at ISPUP, in which projects similar to this one and their procedures were presented, allowing me to better understand about Health Research using statistics.
May	May was essentially dedicated to the structuring and definng this report – organization of results and their interpretation, writing of all procedures and meetings with the goal of monitoring its evolution.
Notes In addition to what was developed as specifically described in each month, during the entire internship, from February to May: - On Thursdays, I attended the weekly class of the "Decision Support Systems in Health" course, taught by Prof. Dr. Milton Severo Barros da Silva, namely on Cluster Analysis (hierarchical methods and non-hierarchical methods) and Classification Models like classification trees, Naive Bayes, Knn, neural networks, etc. - On Mondays, I always had a weekly meeting with Prof. Doctor Milton Severo Barros da Silva, in which we did an update on the work developed in the previous days, new tasks were assigned and the objectives for the week were established.	

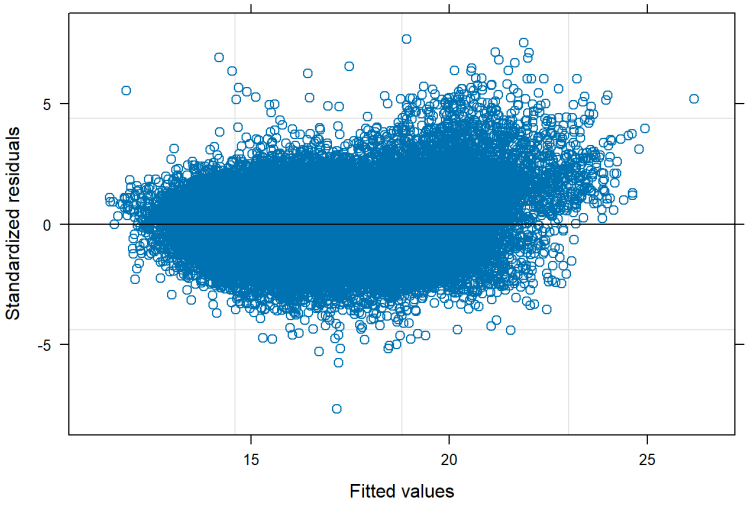
Annex 1: Calendar of Activities of the internship



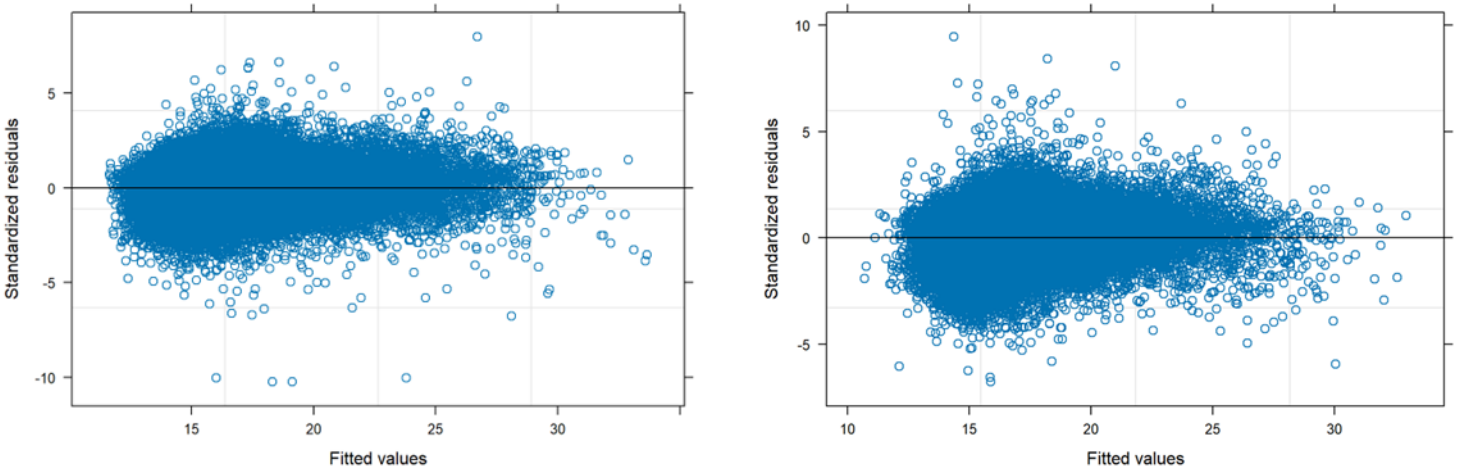
Annex 2: Model fit with the application of the GAMLSS – Females (Source: own elaboration).



Annex 3: Model fit with the application of the GAMLSS – Males (Source: own elaboration).



Annex 4: Data behaviour in model 1 of linear mixed-effects (Source: own elaboration).



Annex 5: Data behaviour in model 2 of linear mixed-effects - Left side with quadratic; Right side with cubic (Source: own elaboration).

FACULDADE DE **ECONOMIA**