ISEG – INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

MASTER IN ACTUARIAL SCIENCE

MASTER'S FINAL WORK

INFLUENCE OF EXTERNAL FACTORS AND FORECAST ON THE CONSUMPTION OF MEDICAL HEALTH INSURANCE

JOÃO MIGUEL ALMEIDA DA CRUZ

SUPERVISORS:

RUTE DE FÁTIMA FIALHO MENDES
PROFESSOR ALFREDO DUARTE EGÍDIO DOS REIS

OCTOBER-2023

Lisbon School of Economics & Management
Universidade de Lisboa

DOCUMENTO PROVISÓRIO ESPECIALMENTE ELABORADO PARA A OBTENÇÃO DO GRAU DE MESTRE

Acknowledgment

This master's thesis would not have been possible without the invaluable support of numerous individuals who have been fundamental in the successful completion of this significant effort. I extend my heartfelt gratitude to all those who have accompanied me throughout this journey.

First and foremost, I express my gratitude to my adviser Professor Alfredo Duarte Egídio Reis, for his diligent supervision, direction, encouragement, and accessibility during the entirety of this undertaking.

Additionally, I would want to express my gratitude to my fellow master's and bachelor's students, in particular to my close bachelor's friend group Mings, and friends who accompanied me during this endeavor, particularly my hometown friends who showed huge support. Their unwavering support and companionship greatly improved my experience, making this journey much more pleasurable and rewarding.

I would also like to thank my team from Médis, who were fundamental for this process, not only for all the help they gave me with this thesis, but also for all the learning and moments I had in the last 6 months. I want to thank particularly to my chiefs Paula Santos and Rute Mendes and also my colleague Raquel Correia for all the help and support and to my friend Miguel Paulino who not only helped with my thesis and work, but also supported me and inspired me during these 6 months.

Finally, and of the greatest importance, I would like to express my sincere gratitude to my parents and my sister for their unwavering support, unwavering encouragement, unwavering friendship, and unwavering patience demonstrated consistently throughout this endeavor. Their invaluable assistance in surmounting the challenges encountered along this path is deserving of my utmost dedication and acknowledgement in this scholarly effort.

Resumo

Este estudo envolveu uma análise estatística abrangente para identificar os elementos subjacentes que influenciam a frequência dos clientes do seguro de saúde Médis em Portugal. O conjunto de dados primário, obtido das bases de dados Sales Index e PorData, consistia numa gama diversificada de variáveis externas, que exigiam pré-processamento e análise abrangentes de dados. Após o processo inicial de categorização, as variáveis foram submetidas a tratamentos particulares que levaram em consideração as suas propriedades temporais e espaciais. Esses tratamentos envolveram a contabilização de fatores como ano, mês e município associados a cada variável. De seguida, foi utilizada a ferramenta analítica Radar para desvendar eventuais correlações que possam existir entre estes fatores e a frequência com que os clientes da Médis interagem com a empresa. Surgiram descobertas importantes, elucidando o impacto de diversas variáveis, como dias da semana, períodos de férias, taxas de poupança, despesas com cuidados de saúde e as ramificações da pandemia da COVID-19. Como resultado, foi tomada a decisão de adotar uma abordagem de GLM em substituição à análise de time series, uma vez que esta última produziu resultados insatisfatórios. A utilização da metodologia do GLM contribuiu para o avanço do nosso entendimento sobre aspetos influentes, fortalecendo o processo de análise e aumentando a nossa capacidade de tomar decisões baseadas em evidências empíricas. A utilização desta técnica abrangente permitiu obter conhecimentos significativos sobre a complexa dinâmica da frequência e consumo dos clientes Médis, melhorando assim os processos de tomada de decisão no setor dos seguros de saúde através do aumento do conhecimento e da eficácia.

Palavras-chave: Time Series, ARIMA, GLM, valor-*p*, previsão

Abstract

This study involved a comprehensive statistical analysis to identify the underlying elements that influence the consume of Médis health insurance clients in Portugal. The primary dataset, obtained from the Sales Index and PorData databases, consisted of a diverse range of external variables, which required comprehensive data preprocessing and analysis. After the initial categorization process, the variables were subjected to particular treatments that took into consideration their temporal and spatial properties. These treatments involved accounting for factors such as the year, month, and municipality associated with each variable. Following that, the analytical tool known as Radar was utilized in order to uncover any correlations that may exist between these factors and the frequency at which Médis clients engage with the company. Prominent findings have surfaced, elucidating the impact of several variables such as weekdays, periods of vacation, rates of savings, healthcare outlays, and the ramifications of the COVID-19 pandemic. As a result, the decision was made to adopt a Generalized Linear Model (GLM) approach as a replacement for time series analysis, as the latter yielded poor results. The utilization of the Generalized Linear Model (GLM) methodology has contributed to the advancement of our comprehension about influential aspects, hence fortifying the modeling process and augmenting our ability to make decisions based on empirical evidence. The utilization of this comprehensive technique has yielded significant insights into the complex dynamics of Médis client frequency, hence enhancing decision-making processes in the healthcare insurance sector through increased knowledge and effectiveness.

Keywords: Time Series, ARIMA, GLM, *p*-value, prediction

Table of Contents

Glossary:

Médis - Companhia Portuguesa de Seguros de Saúde, S.A.: A Portuguese health insurance company.

ISEG: Instituto Superior de Economia e Gestão (Higher Institute of Economics and Management), a Portuguese research and educational institution.

SNS: Sistema Nacional de Saúde (National Health System), the public healthcare system in Portugal.

ADSE: Assistência na Doença aos Servidores civis do Estado (Assistance for Illness to Civil Servants State Employees), a healthcare program for civil servants and state employees in Portugal.

COVID-19: Coronavirus Disease 2019, a global pandemic caused by the SARS-CoV-2 virus.

ARIMA: AutoRegressive Integrated Moving Average, a time series forecasting model.

GLM: Generalized Linear Model, a statistical modeling framework.

AIC: Akaike Information Criterion, a model selection criterion.

ACF: Autocorrelation Function, a tool for measuring autocorrelation in time series data.

BIC: Bayesian Information Criterion, a statistical metric for model selection.

Ljung-Box Test: A statistical test for assessing autocorrelation in time series data.

MAE - Mean Absolute Error: A metric used to measure the average magnitude of prediction errors in regression and forecasting models.

RMSE - Root Mean Square Error: A metric used to measure the square root of the average of squared prediction errors, providing insight into both bias and variance in models.

OLS - Ordinary Least Squares: A method for estimating the parameters of a linear regression model by minimizing the sum of squared residuals.

Figure Index

Chapter 1: Introduction

The current study emerged from a six-month professional internship at Médis - Companhia Portuguesa de Seguros de Saúde, S.A., which constituted an integral component of the master's degree in actuarial science at ISEG. In addition to the application and development of acquired concepts and methodologies, this experience facilitated a deeper understanding of how theoretical knowledge is employed within the actuarial system. Specifically, it provided valuable insights into the analysis and management of data pertaining to each Médis Customer, a significant challenge encountered during the Final Assessment.

This introduction section provides an in-depth description of the methodology utilized in the study to investigate the complex relationship between external variables in Portugal and the frequency of Médis health insurance customers. The study methodology adheres to a sequential approach that involves the gathering of data, meticulous data processing, and meticulous data analysis. The primary aim of this study is to analyze the various aspects that influence the frequency of health insurance clients, utilizing a diverse range of statistical approaches.

The basis of the research is predicated upon a comprehensive procedure for gathering facts. Two key sources of data were employed, both of which served as useful repositories of knowledge. The initial source is the Sales Index database, which is a comprehensive geomarketing tool that encompasses a vast array of socio-economic facts that have been thoroughly documented at the municipal level throughout Portugal. This interactive tool facilitates a comprehensive analysis of the fundamental determinants influencing the frequency of clients in health insurance. The Sales Index database is enhanced by the PorData database, which provides users with access to a diverse range of public data covering both annual and monthly periods. This database comprises data collected from both national and municipal sources, rendering it an essential resource for analysis purposes.

1.1 Data Treatment and Preprocessing

The complexity of raw data typically necessitates the implementation of appropriate treatment and preparation techniques. The original dataset comprised a total of 150,721 observations, which were classified based on the variables of year, month, and municipality. In order to

improve clarity and fully utilize its analytical capabilities, a comprehensive treatment process was implemented. This process entailed additional classification according to the nature of the firm, excluding age (even though this variable has a known impact) as the study wants to exclude every effect from internal variables. Furthermore, careful consideration was devoted to temporal and spatial variables, with a thorough classification of data according to certain time periods (year and month) and geographical regions (municipality).

Nevertheless, as a result of the intrinsic variability of the data, tailored interventions were deemed necessary. The data was classified in a methodical manner according to prescribed rules, taking into account the geographical location (municipality or countrywide) and the temporal aspect (annual or monthly). The outcome yielded a meticulously arranged dataset that is now prepared for subsequent analysis.

After the completion of data processing, the subsequent phase of exploration was initiated, employing the Radar analysis tool. Radar, a tool developed by Willis Towers Watson, is a helpful equipment utilized for the purpose of visual analysis. The utilization of this method enabled the discernment of complex patterns and connections within the dataset, providing insights into the influence of external factors on the frequency of Médis health insurance clients.

The Radar tool provided valuable insights into the impact of various variables, including the number of working days, school vacations, household savings rates, expenses of the Sistema Nacional de Saúde (SNS) and Assistência na Doença aos Servidores Civis do Estado (ADSE), and the significant influence of the COVID-19 pandemic on the frequency of health insurance clients. The visualizations shown in this study served as a guide for future analytical procedures.

1.2 Time Series Analysis

In order to obtain more profound insights and make predictions about future trends, a methodology known as Time Series Analysis was employed. This approach entailed the utilization of the AutoRegressive Integrated Moving Average (ARIMA) model. The process encompassed several key components, namely data partitioning, model estimation, forecasting, assessment of model performance, and diagnostic examinations. The dataset was partitioned into two sets: a training set, which comprised 80% of the data, and a test set, which comprised

the remaining 20%. This division allowed the model to be trained using past data and subsequently assessed for its ability to make predictions on new, unknown data.

The ARIMA(3,1,1) model, characterized by its autoregressive, differencing, and moving average components, provides a robust framework for examining temporal interdependence. The utilization of diagnostic tests facilitated the identification of areas requiring improvement, therefore pinpointing specific instances in which the model's predictions diverged from the observed data.

1.3 Generalized Linear Model

Despite conducting a comprehensive time series analysis, the obtained results did not align with the expected outcomes. In order to achieve a more comprehensive understanding, a transition was made towards employing a Generalized Linear Model (GLM). This transition was made for a more in-depth examination of the 51 predictor variables and their influence on the frequency of Médis health insurance clients.

The Generalized Linear Models (GLM) framework has been found to improve the interpretability and analytical efficiency of statistical analyses, hence facilitating the investigation of variable relationships in a more simplified manner. The study utilized Gaussian family regression, capitalizing on the visual indicators of a normal distribution present in the dataset. This methodology enabled the determination of the amount and direction of impact for each independent variable. Variable selection was led by statistical significance, which was tested using p-values, so aiding in the refinement of the model.

Throughout the course of the study, ongoing evaluation of the prediction capacities of the models was carried out. The evaluation of model predictions is a crucial component of the study. The normal distribution of predictions plays a crucial role in suggesting assessments that are unbiased and well-calibrated. A range of statistical tests and visualizations, including histograms, were employed to assess the normality of predictions, so enhancing the level of confidence in the efficacy of the model.

This detailed overview provides a complete analysis of the research conducted to understand the complex correlation between external circumstances in Portugal and the frequency of Médis health insurance clients. The methodology employed in this study is marked by a rigorous approach to data treatment, thorough exploratory analysis, the use of time series

modeling techniques, and the transition to Generalized Linear Models. These methodological choices reflect a strong dedication to producing reliable and informative conclusions. The forthcoming chapters will explore the outcomes of this research, providing insights into the determinants influencing the frequency of health insurance clients in Portugal.

Chapter 2: Literature Review

2.1 The impact of health insurance and external variables

Health insurance functions as a protective mechanism, providing individuals with financial covering for their medical expenses and facilitating their access to vital healthcare services, thereby alleviating the full extent of financial responsibility. In the usual course of events, individuals make regular payments to their selected insurance provider, and as a result, the insurer assumes partial or complete responsibility for their healthcare costs, which include appointments with medical professionals, hospital accommodations, prescribed drugs, and measures taken to prevent illness or injury. The influence of health insurance on the broader populace is significant, exerting a dramatic impact on overall health and welfare (Acharya et al., 2013).

The impact of health insurance on the wider populace is significant. The primary objective of this initiative is to enhance health outcomes by removing the financial obstacles that might discourage people from accessing medical services. The presence of insurance coverage increases the likelihood of individuals participating in regular check-ups, swiftly addressing illnesses, and effectively managing chronic disorders. The implementation of a proactive healthcare approach not only serves to improve the health of individuals, but also has a positive impact on the overall well-being of the population. This is achieved through the reduction of communicable disease transmission and the containment of long-term healthcare expenses. Moreover, health insurance engenders a feeling of assurance, mitigating the anxiety and economic burden associated with unforeseen medical crises, thus augmenting psychological welfare and overall standard of living. Health insurance continues to be a crucial instrument in promoting a society that is both healthier and more resilient (Acharya et al., 2013).

Some of the variables that influence this study were previously shown to be significant in matters of health of the public. This study selected variables based on prior research. An example of previous research study relates to the examination of certain variables, with particular emphasis on the consequences of the lockdown measures imposed as a result of the pandemic (Moynihan et al., 2020) where the study says that were “reported 143 estimates of changes in healthcare utilization between pandemic and pre-pandemic periods”. An additional instance of prior research refers to a study that explains the impact of domestic water quantity.

This investigation prompted the inclusion of a variable representing the percentage of consumable water in a given municipality (Howard et al., 2003). Furthermore, the study incorporated several environmental variables, such as the proportion of land area affected by fire, within a certain municipality (Cascio, W. E., 2018) and the quantity of urban residuals from municipalities (Beebe, A., 2021).

There were some social and economy-based variables incorporated on the study such as inflation and inflation on healthcare (Salatin, P., & Bidari, M., 2014) or saving rates per family (Białowolski et al., 2019) being examples of this type of variables.

2.2 Regression

Regression refers to a class of mathematical and statistical methods employed to establish a model that captures the association between a dependent variable and a collection of independent variables. The nature of this interaction can be intricate, giving rise to the emergence of diverse methodologies within this conceptual framework. Regression analysis can be utilized for two distinct categories of data, namely continuous data, or discrete data. When confronted with a continuous dependent variable, it is customary to employ the normal distribution as a model for its distribution. On the other hand, generalized linear models are utilized in cases when the dependent variable is not of a continuous nature. Hence, regression analysis can be employed for analyzing both continuous and discrete data. It is worth noting that regression methods for continuous data are typically more straightforward in comparison to those specifically developed for discrete data.

2.2.1 Linear Regression

The linear model, as expounded upon by (Neter, Kutner, Nachtsheim, Wasserman et al. 1996), is a generally acknowledged statistical paradigm that seeks to build a correlation between a dependent variable, which adheres to a normal distribution, and a set of independent variables. The thesis posits that the dependent variable can be represented as a linear function of the independent variables.

Consider a hypothetical dataset of significant magnitude consisting of n rows and $p-1$ columns. This dataset represents the existence of n samples or observations, each possessing

p−1 distinct properties. In a broader perspective, the representation of a linear model can be mathematically stated as described in equation (2.1).

$$y = X\beta + \varepsilon \qquad (2.1)$$

where y is a vector of dependent variables X is the independent variables, β is the vector of parameters and ε is the vector of errors.

2.2.2 Generalized Linear Model

The Generalized Linear Model (GLM) is commonly acknowledged as a versatile extension of the traditional Ordinary Least Squares (OLS) regression. This extensive classification of models contains both conventional regression and ANOVA models is common in the analysis of continuous responses. Additionally, there exist models that are specifically tailored for the analysis of categorical responses. GLMs consist of three fundamental components. (Agresti, 2007).

The initial component in the analysis pertains to the random element, which serves as the response variable Y and establishes the presumed distribution for said variable.

The second component, known as the systematic component, entails the determination of the explanatory variables employed in the model.

The third component of the model is the link function, represented as g(·), which acts as the intermediary between the random and systematic components. The function in question serves to construct links, whether linear or nonlinear, between a collection of fitted values derived from the model and the predictor variables.

The selection of a link function determines the precise structure of the model, which can be conventional regression $g(\mu) = \mu$, loglinear $g(\mu) = \log(\mu)$, or logistic regression $g(\mu) = \log(\frac{\mu}{1-\mu})$. The chosen model for this study is linear regression, which is highly acknowledged and renowned for its applicability across numerous applications.

2.2.3 *p*-value and Tests

The p-value holds significant importance as a statistical indicator for assessing the significance of predictor variables within the framework of generalized linear models (GLMs). The primary objective of this indicator is to provide a quantitative evaluation of the evidence opposing the null hypothesis (Dahiru, T., 2008). The null hypothesis suggests that there is no statistically significant link between a specific predictor variable and the response variable. During the process of fitting a Generalized Linear Model (GLM), a p-value is assigned to each predictor variable. The calculation of the p-value is derived from the estimated coefficients and their associated standard errors. A p-value with a low value, typically determined by a predetermined significance level like 0.05, offers significant support for the rejection of the null hypothesis. This discovery suggests that the predictor variable makes a considerable contribution to explaining the variability observed in the responder variable (Dahiru, T., 2008). Conversely, a high p-value suggests insufficient evidence to reject the null hypothesis, suggesting that the predictor variable may lack a substantial influence on the model. Researchers utilize p-values to make informed decisions regarding the incorporation or removal of variables within Generalized Linear Models (GLMs). This methodology contributes to the enhancement of the model by prioritizing the importance of predictors that have attained statistical significance (Dahiru, T., 2008).

The Akaike Information Criterion (AIC) provides an alternate approach to model evaluation in generalized linear models, as opposed to the traditional use of p-values. The Akaike Information Criterion (AIC) functions as a comprehensive metric for evaluating the overall quality of a model. The objective is to achieve an appropriate balance between the model's level of accuracy in representing the data and its level of intricacy. (Bozdogan, H., 1987). The computation of this method is dependent on the log-likelihood of the model and the number of parameters it utilizes. Models with lower AIC values are indicative of better fits, as they demonstrate the ability to accurately reflect the underlying characteristics of the observed data while also preserving a level of simplicity. Researchers often rely on the Akaike Information Criterion (AIC) as a valuable tool for model selection when faced with numerous potential models (Bozdogan, H., 1987). The AIC serves as a guiding compass, aiding researchers in making informed decisions regarding the most appropriate model to choose. The approach promotes the choice of models that achieve an optimal equilibrium, allowing them to

properly depict the data without falling prey to overfitting. The AIC improves the modeling process by giving priority to the principles of precision and quality of fit.

2.3 Time Series

2.3.1 Time Series Origin and Models

The development of time series analysis can be traced to the early 20th century, characterized by the groundbreaking contributions of Yule as evidenced in a collection of seminal works (Yule, 1921, 1926, 1927).

Yule (1921) conducted a study in which he aimed to investigate the phenomenon of time-correlation, specifically exploring the associations between unrelated variables observed across different time periods.

In 1926, Yule's focus moved to examining the correlations between variables related to time, which he humorously referred to as "nonsense-correlations." This study not only provided insights into these correlations but also uncovered the complex association between serial correlations within a series (ri) and those within the difference series (ρi).

Yule's seminal research expanded beyond its initial scope. The individual in question established the theoretical framework of autoregressive series and proceeded to employ a specific type of autoregressive process, namely the second-order autoregressive process denoted as AR(2), in order to construct a model for Wolfer's sunspot data. This dataset consists of consecutive annual measurements of sunspot numbers. The aforementioned study demonstrated the potential of utilizing previous observations of a variable to gain understanding of and exert influence on its current behavior (Yule, 1927).

The second-order autoregressive process proposed by Yule, commonly written as AR(2), is defined by a specific equation that can be represented as

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t \qquad (2.2)$$

Within this particular framework, the variable y_t signifies the observation at a distinct point in time, but the coefficients a_1 and a_2 are employed to symbolize the regression coefficients, also known as autocorrelations. Furthermore, it is postulated that ε_t denotes uncorrelated stochastic disturbances with a mean of zero and a variance of one.

Yule's first research established the foundational principles, which were subsequently built upon to develop a more comprehensive autoregressive model referred to as AR(p). In Walker's seminal study published in 1931 (Walker, G. T., 1931), it was empirically established that the model under investigation successfully represents the interdependence between successive, unbroken terms within the series denoted as y_t. Furthermore, the model takes into account the existence of an error term denoted as ε_t. The model is referred to as

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + \varepsilon_t \qquad (2.3)$$

The study (Walker, G. T., 1931) demonstrated that, with a sufficiently large number of observations (n), an equivalent equation may be derived. The aforementioned equation, however, does not incorporate the stochastic disturbance ε_t, and it establishes a connection between the sequential correlation coefficient values (often referred to as autocorrelations) of the elements inside the series, r_i.

$$r_k = a_1 r_{k-1} + a_2 r_{k-2} + \ldots + a_p r_{k-p} \qquad (2.4)$$

The Yule-Walker equations, sometimes denoted as a system of linear equations, are defined for values of k ranging from 1 to p. The notion of moving average (MA) models was first established by Slutsky (Slutsky, E., 1937), with a focus on exploring the potential association between these models and cyclical processes.

Wold made significant advancements in the field of data analysis by introducing the utilization of moving average processes (Wold, H., 1938). The study conducted by the author demonstrates the possibility of decomposing a stationary time series into a moving average model consisting of independent random variables. Nevertheless, the utilization of moving average models was constrained as a result of the difficulties involved with identifying appropriate models and the absence of efficient techniques for model selection, fitting, and evaluation.

In order to tackle these issues and offer increased adaptability in the modeling of real-time series data, the integration of moving average (MA) and autoregressive (AR) processes was undertaken, leading to the emergence of autoregressive moving average (ARMA) processes.

The mathematical and statistical ideas that form the basis of Autoregressive (AR), Moving Average (MA), and Autoregressive Moving Average (ARMA) processes can be studied in the publications of Box and Jenkins (1970, 1976), as well as Box et al. (1994, 2008). It is crucial to underscore that ARMA models are specifically designed for the examination of stationary time series data.

Box and Jenkins expanded their inquiry beyond time series that exhibit stationarity by employing mathematical statistics and probability theory (Box, G. E. P. and Jenkins, G. M., 1970). The researchers expanded the range of ARMA models to include particular categories of nonstationary time series, so developing a group of models known as autoregressive integrated moving average (ARIMA) models. The ARMA and ARIMA models have been acknowledged as effective tools with broad applicability to several academic disciplines for analyzing time series data.

Moreover, Box and Jenkins expounded upon a coherent and pragmatic three-step iterative methodology for the purpose of time series modeling (Box, G. E. P. and Jenkins, G. M., 1970). This methodology comprises the processes of model identification, parameter estimation, and model validation. The commonly accepted approaches, such as the Box-Jenkins technique, facilitate the estimation of crucial parameters such as p, d, q, and others within an appropriate ARIMA(p, d, q) model for a given dataset. In this context, the variables p, d, and q denote non-negative integers that serve as indicators for the orders of the autoregressive, integrated, and moving average components of the model, respectively.

2.3.2 Tests in Time Series Analysis

The Autocorrelation Function (ACF) holds a fundamental position within the field of time series analysis. The proposed method aims to measure the magnitude of linear associations between data points within a time series and their corresponding values at different time lags (Hansen, P. R., & Lunde, A., 2014). The Autocorrelation Function (ACF) provides insights into the relationship between each observation in a time series and its preceding values. This tool possesses significant value in the evaluation of serial correlation, which pertains to the correlation between a data point and its past observations.

When analyzing an autocorrelation function (ACF) plot, researchers scrutinize notable peaks or patterns that surpass the bounds of the confidence interval. The presence of a spike at a particular lag indicates a robust correlation between data points that are separated by said lag.

An ACF plot that exhibits a progressive decrease as the lag increases indicates a correlation between earlier observations at various lags and the current observation (Hansen, P. R., & Lunde, A., 2014). This observation indicates the presence of a non-stationary time series, which may require the application of differencing in order to attain stationarity. On the other hand, the presence of a distinct and abrupt decline in the autocorrelation function (ACF) plot following a small number of lags suggests the presence of stationarity (Hansen, P. R., & Lunde, A., 2014). This characteristic facilitates the process of selecting a suitable AutoRegressive Moving Average (ARMA) model for the given time series.

The Bayesian Information Criterion (BIC) is a commonly adopted statistical metric in the field of time series analysis. It is utilized for model selection and hypothesis testing purposes. The discussed notion attains a balanced equilibrium between the accuracy of a statistical model in precisely describing the data and the complexity inherent in the model. The Bayesian Information Criterion (BIC) integrates two fundamental components: the likelihood of the observed data given the model, and the number of parameters utilized in the model (Neath, A. A., & Cavanaugh, J. E., 2012). The primary objective of this methodology is to discourage the use of overly complex models, since they have the potential to result in overfitting, a scenario in which the model excessively adapts to the provided data and subsequently fails to effectively generalize to new data instances. The Bayesian Information Criterion (BIC) assigns greater scores to models that demonstrate a robust fit to the observed data, while also encouraging parsimony by penalizing models with an excessive number of parameters. Academics frequently utilize the Bayesian Information Criterion (BIC) to assess and contrast different models, with the objective of determining the model that achieves a harmonious equilibrium between the quality of fit and the complexity of the model (Neath, A. A., & Cavanaugh, J. E., 2012). This methodology improves the process of making informed decisions in various fields, such as statistics, machine learning, and economics.

The final test investigated in this research was the Ljung-Box test, which carries substantial significance as a statistical methodology within the realm of time series analysis. The primary objective of this tool is to assess the existence of significant autocorrelation in a time series, hence aiding in the detection of probable interrelationships or underlying patterns within the observed data points (Hassani, H., & Yeganegi, M. R., 2019). The methodology involves the computation of a test statistic by utilizing autocorrelation coefficients at different lag intervals, followed by a comparison of this statistic with a critical value derived from the

chi-square distribution. When the computed test statistic exceeds the critical value, it signifies the presence of autocorrelation within the time series data, suggesting that the data points are not independent. Conversely, if the test statistic is less than the critical value, it indicates that the time series can be effectively characterized as a white noise process, wherein the data points demonstrate no association (Hassani, H., & Yeganegi, M. R., 2019). The Ljung-Box test is a crucial method employed in evaluating the suitability of time series models and confirming adherence to the concepts of independence and randomness.

Chapter 3: Methodology

3.1 Treatment and Analysis of the data

The data related to external factors in Portugal were initially obtained from the Sales Index database, which is a geomarketing application that incorporates the primary socio-economic statistics accessible in Portugal at the municipal level. This database facilitates various value-added analyses for the development of sustainable geomarketing strategies. Additionally, the PorData database was also used in this process. The statistics presented in this study were obtained by extracting information from public sources, including annual and monthly data, as well as data at both national and municipal levels. A comprehensive set of 51 variables was derived from the previously mentioned approach and further subjected to correlation analysis with Médis data based (data extracted from Médis database) on temporal factors such as year and month, as well as spatial factors such as municipality. The table including the variables utilized in the method is shown in Appendix I.

The initial dataset contains 150,721 observations prior to data treatment. As previously mentioned, these variables were initially categorized based on year, month, and municipality. However, a subsequent request was made to further categorize the data based on the type of business (e.g., Individual, PMEs, etc.). It is imperative to note that the timeline employed in the procedure spans from 2017 to 2022.

Due to the varying nature of several variables, it was necessary to apply specific treatments to each variable excluding age (even though this variable has a known impact) as the study wants to exclude every effect from internal variables, based on predefined guidelines. The set of regulations contained a predetermined set of guidelines.

i. Is the data categorized by municipality or on a nationwide level? In the case when the variable pertains to a specific municipality, it should be sent to rule number ii). On the other hand, if the variable is of a national nature, it should be directed to rule number iii).

ii. The following question related to the temporal nature of the variable, namely whether it was measured on an annual or monthly basis. If the variable was already recorded on a

monthly basis, no modifications were necessary, and it could be assigned to the Médis datamart. If the variable is based on an annual basis, rule number iv) is applied.

iii. If the variable were to be considered at a national level, a similar question arises as to whether the information pertains to the entire year or if it is based on a monthly basis. The variable is allocated to the Médis datamart on a monthly basis, specifically by municipality. If the variable is based on an annual basis, rule number v) is applied.

iv. Given that the variable under consideration is of an annual nature, the determination was reached by assessing its divisibility into individual months (for example the number of fires in each municipality was divided into every month) or if the variable contains only annual information that cannot be disaggregated into monthly data, the information is assigned to the datamart variable as it was (an example of this is the number of health centers on a municipality).

v. In an approach similar to the previous point iv), the identical procedure is applied to variables based on national-level data, afterwards assigning them to the datamart on a yearly basis.

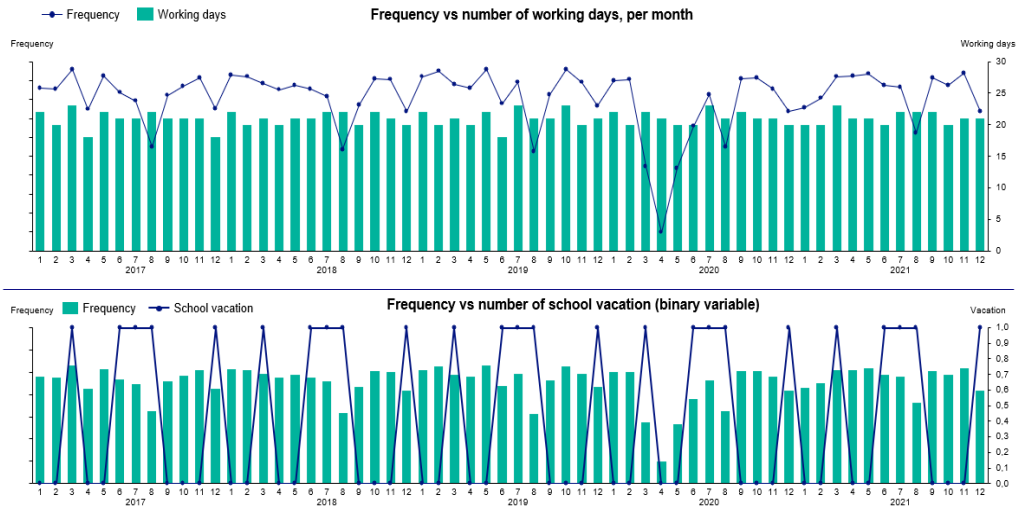
vi. Lastly, if a variable is contingent upon the municipality, adjustments were made to the variable based on the specific municipality, taking into account the nature of the variable (for example the burned area was divided by the total area of a municipality to get the percentage of burned area per municipality).

Following the initial data processing, an analytical tool known as Radar, developed by Willis Towers Watson, is employed to visually examine the potential impact of variables on the frequency of Médis health insurance clients. Radar is a software application designed to interpret user-provided data and facilitate graphical representation of data trends. The frequency was determined at Radar by dividing the number of payable claims (the amount an insurer is obligated to pay to a policyholder following a valid claim) by the total exposure (the level of risk an insurer faces from its policyholders within a specific period) per insured individual in a given calendar year.

Following the creation of the images on Radar, it was observed that certain patterns exist between the frequency and select variables examined in the study. The subsequent section will go into the exploration of these identified correlations.

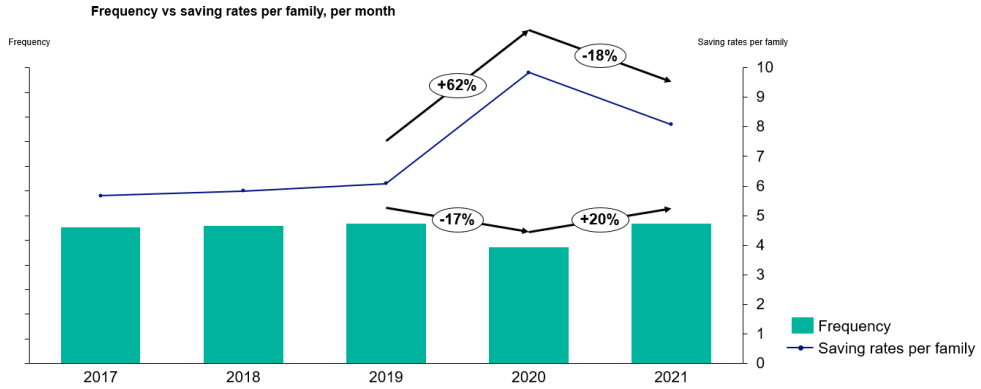
The initial finding pertained to the relationship between the frequency and the quantity of working days and vacations. Upon analyzing the graphical representation (Figure 1), it becomes evident that the frequency exhibits an upward trend during months characterized by a greater number of working days, with the exception of those months wherein the duration of school vacation holds substantial influence. In the months of July and August, it is seen that the number of working days is somewhat larger compared to other months of the year. However, the usage frequency of the Médis portfolio is lower during this period due to the major impact of school vacations.

Figure 1 – Graphics comparing the frequency to working days and vacation



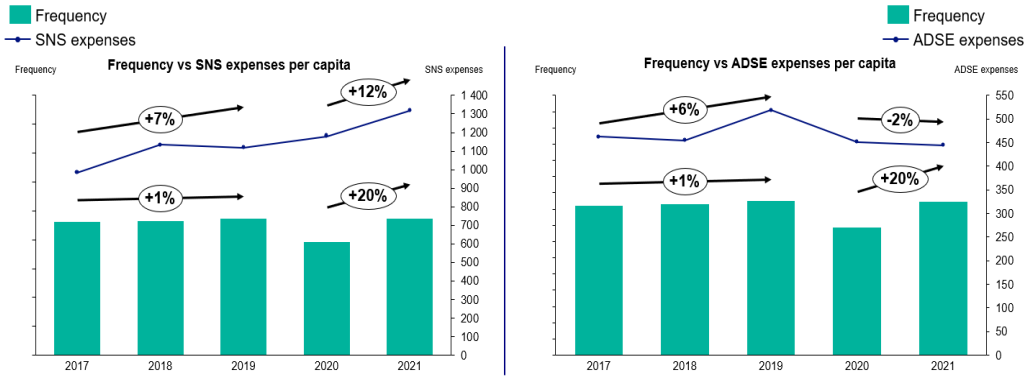
Following that, an observation was conducted regarding the savings rates per household. The impact of the COVID-19 pandemic in 2020 has significantly impacted the relationship between frequency and saving rates per family. However, it is worth noting that from 2017 to 2019, there was a tendency for saving rates per family to increase as frequency increased. In the years 2020 and 2021, there has been a notable shift in behavior, as evidenced by the data presented in the graph (Figure 2). Specifically, there has been a significant increase of 62% in saving rates per family as a result of the pandemic. However, it is important to note that these savings have subsequently experienced a decline of 18%. Conversely, there has been a growth of 20% in the frequency of certain activities.

Figure 2 – Comparison between frequency per year and saving rates per family



Subsequently, a graphical analysis was conducted to observe the expenses per capita of the Sistema Nacional de Saúde (SNS) and the Assistência na Doença aos Servidores Civis do Estado (ADSE). This data reveals an upward trend in both SNS and ADSE expenses, as well as an increase in frequency, between the years 2017 and 2019. In the year 2020, there was a notable decline in both frequency and ADSE, despite the continued growth in SNS expenditures. In the year 2021, there has been a notable increase in both the frequency and expenses associated with social networking services (SNS) when compared to the previous year, 2020. This growth can be attributed to the recovery from the effects of the pandemic. However, it is worth noting that the expenses related to the acquisition, development, and support of enterprise software (ADSE) have continued to decline, albeit not as significantly as observed in the year 2020 in contrast to the year 2019.

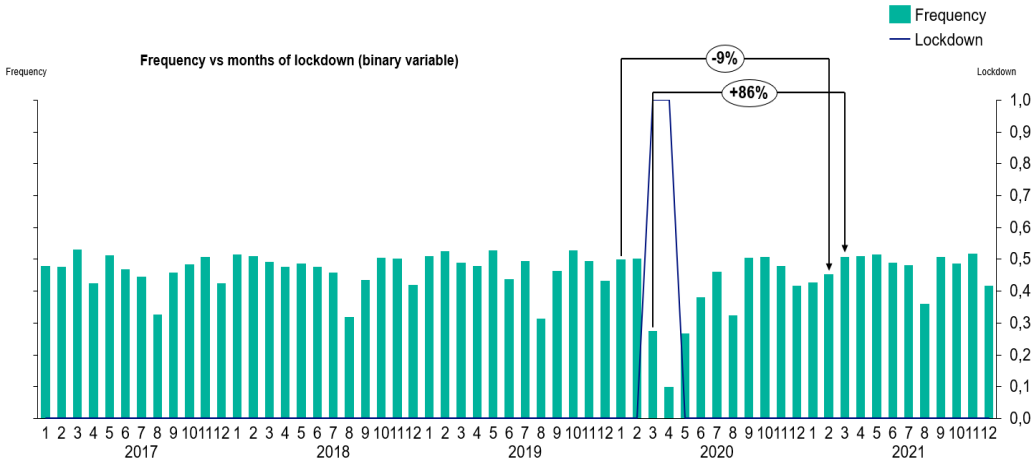
Figure 3 – Comparison between frequency and SNS and ADSE annual expenses



Finally, a crucial aspect of the analysis is the examination of the pandemic's frequency, given its significant influence on the data. The chart illustrates a notable decline during the months of March and April in the year 2020, which can be attributed to the implementation of

a lockdown as a result of the COVID-19 pandemic. The observed decline, relative to the previous year, resulted in a 9% decrease in frequency from January 2020 to January 2021. However, from March 2020 (the onset of the pandemic and subsequent lockdown measures) to March 2021, there was an 86% increase in frequency. This suggests a recovery in frequency, likely attributable to the impact of the pandemic and the adjustment of frequency levels to pre-pandemic conditions.

Figure 4 – Comparison between Frequency and the months affected by lockdown



Following the completion of the exploratory analysis using Radar, it became necessary to implement data modification. The data was acquired from the Sales Index and PorData, as mentioned earlier. However, several variables lacked the necessary information to conduct the models. The software employed for making alterations to the data and implementing the model was RStudio.

The initial observation made regarding the data pertained to the absence of revealed municipality information for some data points. Consequently, these cases were excluded from the analysis. Subsequently, it was seen that certain minor towns incorporate a year or a few years' worth of omitted data. The approach employed to address the matter of small municipalities was establishing a mechanism for obtaining data from the subsequent year. This was done due to the absence of any instances where information was unavailable for the years 2021 and 2022. Consequently, data from the year preceding this was incorporated.

Following the previously mentioned process of incorporating the previous year, it was observed that certain minor municipalities in subsequent years exhibited a dearth of data pertaining to the variable associated with business type. Consequently, it was determined that

the inclusion of business type in the study would be omitted. The reasoning behind this scenario is because in cases when data related to a specific variable was unavailable for a particular year, and during that year a new business type was introduced to the municipality, it became difficult to determine the preceding value or make adjustments to that value.

The last alteration was intended for municipalities that lacked any available data across all the years. Given that the municipalities under consideration were rather small, the missing values were replaced with the mode of the variables that were missing. This approach was considered appropriate due to the larger number of small municipalities in Portugal and in the study, which lends greater credibility to the chosen value.

3.2 Time Series Analysis

Following the data treatment and elimination of missing values, a time series analysis was conducted to estimate forecast values based on the frequency. The research utilizes an ARIMA (AutoRegressive Integrated Moving Average) model to effectively capture the temporal relationships that exist within the time series data. The primary stages are the division of data, the fitting of models, the prediction of outcomes, the evaluation of performance using error metrics, and the utilization of diagnostic tests to evaluate the suitability of the selected model. The subject matter provides a comprehensive explanation of every individual stage.

The initial procedure involves partitioning the dataset into two distinct subsets: a training set, encompassing 80% of the data, and a test set, encompassing the remaining 20%. The separation of data provides a crucial function as it allows us to train the model using past data and afterwards evaluate its predictive capabilities on novel, unseen data. The training set furnishes the model with the requisite information for comprehending and identifying patterns and correlations present in the dataset.

The primary focus of the analysis revolves around the utilization of an ARIMA(3,1,1) model on the training data. The model definition comprises three fundamental components: autoregressive (AR) terms, differencing (I), and moving average (MA) terms. The autoregressive (AR) component of the model incorporates the notion of how previous observations impact the current value. The inclusion of the differencing component in the model signifies that the data is subjected to a single differencing operation in order to attain stationarity, which is a necessary condition for the application of ARIMA modeling. The inclusion of earlier forecast errors in the current projection is shown by the MA component.

The ARIMA model that has been fitted quantifies temporal dependencies and offers coefficients for the autoregressive (AR) and moving average (MA) terms.

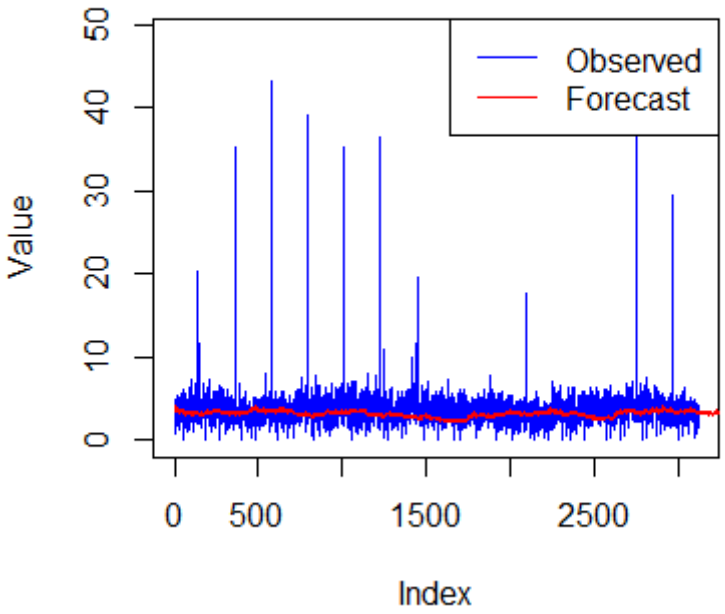
Figure 5 – RStudio code that estimates the Time Series model

```
arima_model_final <- auto.arima(train_data$Frequencia, seasonal = TRUE, stepwise = TRUE,
                                approximation = FALSE, seasonal.test = "ocsb");sarima_model
```

The code sample presented uses the auto.arima() method to estimate an ARIMA (AutoRegressive Integrated Moving Average) model using the training data, specifically the variable Frequencia. This process is guided by several critical criteria. To begin, when the seasonal parameter is set to TRUE, the function integrates seasonal patterns present in the data, which may occur at regular intervals such as daily, weekly, or yearly, into the ARIMA model. Furthermore, the inclusion of the "stepwise = TRUE" parameter triggers a stepwise search procedure aimed at identifying the most suitable ARIMA model. This is accomplished through a systematic exploration of different parameter combinations, ultimately picking the model with the lowest AIC (Akaike Information Criterion) value. The inclusion of the "approximation = FALSE" option guarantees the utilization of a precise technique for estimating the ARIMA model, which may yield more precise outcomes, but at the cost of additional computational requirements. The parameter "seasonal.test = "ocsb" is used to indicate the utilization of the "OCSB Test" (Osborn, Chui, Smith, and Birchenhall Test) as the chosen approach for assessing seasonal trends in the time series data.

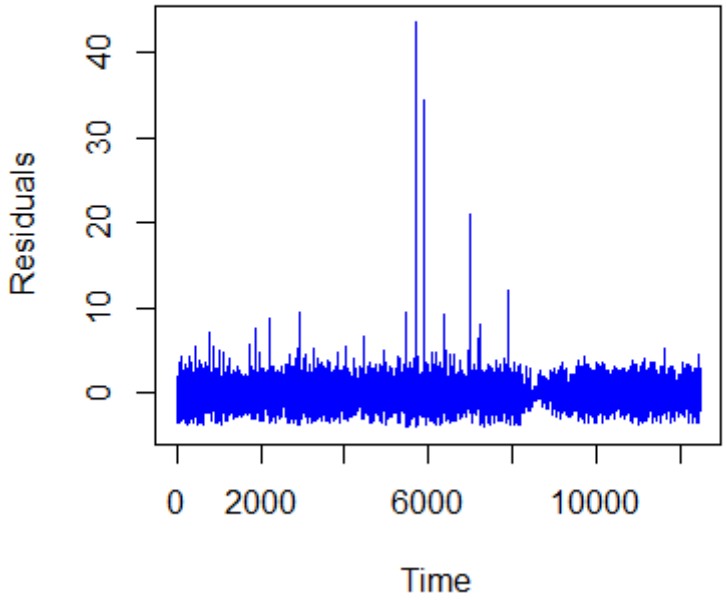
The ARIMA model's results include confidence intervals that are shown alongside the anticipated values. The terms "Lo 80" and "Hi 80" refer to the lowest and upper limits of an 80% prediction interval, respectively. Similarly, "Lo 95" and "Hi 95" indicate the corresponding boundaries for a 95% prediction range. These intervals include the whole spectrum of potential outcomes for each projected number. As an example, consider the projected value at time point 12469, which is estimated to be roughly 3.378358. This estimation is accompanied by an 80% prediction interval ranging from 1.499118 to 5.257598. This suggests that, with a confidence level of 80%, it is anticipated that the true value will be within this range.

Figure 6 – Graphic comparing the observed data and the forecasted data



In order to assess the prediction performance of the model, two key error metrics are calculated: the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). The Mean Absolute Error (MAE) provides a quantitative assessment of the average magnitude of prediction mistakes, hence offering valuable information regarding the model's inherent bias. In this specific case, an estimation of the Mean Absolute Error (MAE) yields an approximate value of 1.086. In addition, the Root Mean Square Error (RMSE) enhances the assessment by considering both the systematic deviation and the variability of errors. It is computed as the square root of the average of the squared errors. Currently, the estimation of the root mean square error (RMSE) stands at approximately 1.466. These measures provide useful insights into the performance of the model, enabling comparisons with competing models or forecasting methodologies.

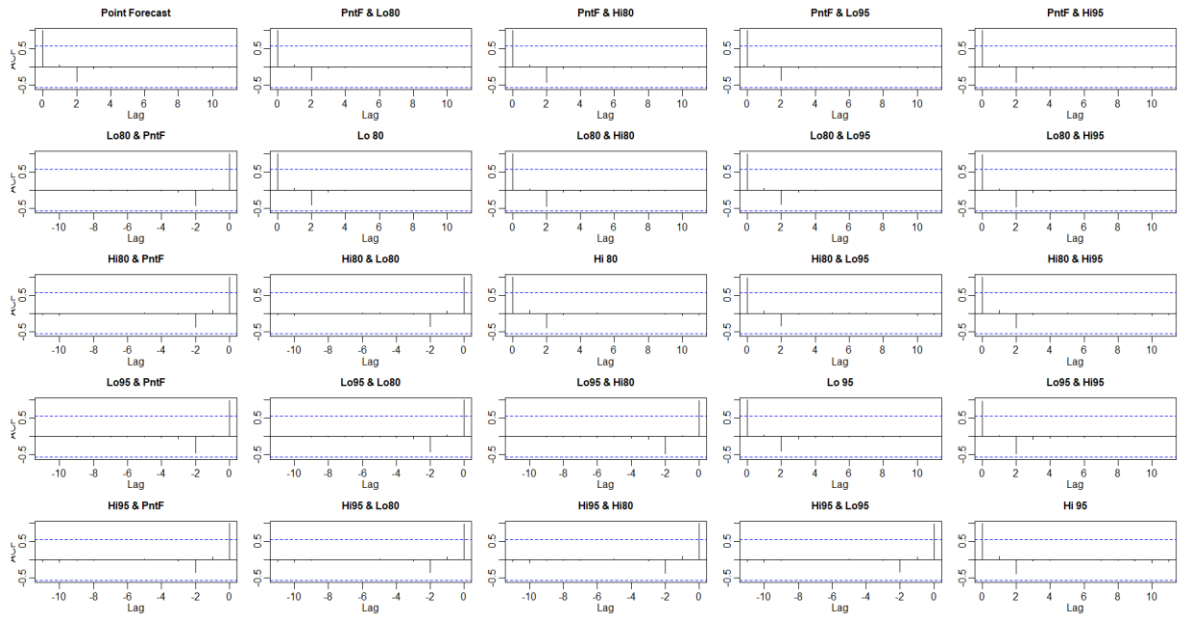
Figure 7 – Display of the residuals



Upon examination of the residual plot, it becomes evident that a discernible pattern arises, characterized by the majority of data points clustering around four unique peaks along the y-axis. These peaks indicate occurrences where the residuals exhibit significant departures from zero. These deviations may indicate certain time points or intervals during which the model's predictions exhibited substantial divergence from the observed data. However, a notable feature of this plot is that, aside from the four obvious peaks, the remaining residuals exhibit a continuous and stable alignment along the y-axis. This suggests that, on the whole, the model successfully represents the fundamental patterns and trends present in the data, except for the noticeable spikes. In order to assess the predicted accuracy of the model, two key error metrics are calculated: the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). The Mean Absolute Error (MAE) provides a quantitative assessment of the average magnitude of prediction mistakes, hence offering valuable information regarding the model's bias. In the present situation, the Mean Absolute Error (MAE) is approximately calculated to be 1.086, as previously stated. In addition, the Root Mean Square Error (RMSE) enhances the evaluation process by considering both the systematic deviation and the spread of mistakes. It is computed as the square root of the average of the squared errors. In the current context, the root mean square error (RMSE) has been previously reported to be approximately 1.466. These measures

provide useful insights into the performance of the model, allowing comparisons with other models or forecasting methodologies.

Figure 8 – ACF visual representation



The AutoCorrelation Function (ACF) plot is a visual representation used to illustrate autocorrelation patterns. It provides insights into the relationship between a particular data point in a time series and data points at various delays or time intervals. The horizontal axis of the graphic indicates different lag levels, while the vertical axis displays the corresponding autocorrelation coefficients.

The ACF plot displays the "Point Forecast" which reveals the predicted values at various lags. It is worth mentioning that the autocorrelation tends to converge to a value of 1 at lag 0, which signifies a robust positive association between a given data point and itself in the absence of any time lag. In contrast, when considering a lag of 1, the autocorrelation approaches zero, indicating that there is little to no correlation between nearby data values at consecutive time periods. At a lag of 2, the autocorrelation exhibits a value of around -0.5, suggesting a moderate negative correlation between data points that are separated by a time interval of 2.

Furthermore, the ACF plot incorporates "PnF & Lo80" and "PnF & Hi80" to denote the prediction intervals linked to the point forecasts. The acronym "PnF" is an abbreviation for "Point Forecast," which denotes the projected value at a particular time lag. The terms "Lo80" and "Hi80" are used to represent the bottom and upper limits of the 80% prediction interval, respectively. These intervals establish a range in which the true values are anticipated to lie

with a confidence level of 80%. In a similar vein, the terms "Lo95" and "Hi95" are used to denote the lower and upper limits of the broader 95% prediction interval, which provides a heightened degree of confidence.

The ACF plot provides a comprehensive visual representation of autocorrelation coefficients at different delays. Positive autocorrelation values show a positive association between data points at various lags, whilst negative values indicate a negative association.

The utilization of this complete autocorrelation function (ACF) plot aids in the evaluation of the presence of statistically significant autocorrelations within the residuals of the model. In an optimally fitted model, the statistical significance of autocorrelation values at various lags should be negligible, suggesting that the model successfully represents the intrinsic temporal dependencies present in the data.

To summarize, the code segment provided utilizes the auto.arima() function to fit an ARIMA model. The ACF plot and its components are used to assess the autocorrelation structure present in the residuals. These insights are of utmost importance in assessing the extent to which the model effectively incorporates the temporal dependencies inherent in the time series data. Upon examination of the residual plot, it becomes evident that a discernible pattern is present. The majority of data points exhibit a tendency to cluster around four distinct peaks along the y-axis. These peaks indicate points at which the residuals exhibit significant departures from zero. These deviations may suggest particular time points or intervals during which the model's predictions deviated significantly from the actual observations. However, a notable observation in this plot is that, with the exception of these four noticeable peaks, the residual values exhibit a rather uniform and constant distribution down the y-axis. This suggests that, on the whole, the model successfully represents the fundamental patterns and trends present in the data, save for the noticeable peaks. In order to assess the prediction performance of the model, two key error metrics are calculated: the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). The Mean Absolute Error (MAE) serves as a metric for quantifying the average magnitude of prediction mistakes, hence offering valuable insights on the bias exhibited by the model. In this particular instance, the Mean Absolute Error (MAE) is estimated to be roughly 1.086. In addition, the Root Mean Square Error (RMSE) enhances the evaluation process by considering both the systematic deviation and the variability of errors. It is computed as the square root of the average of the squared errors. In the present situation, the root mean square error (RMSE) is estimated to be roughly 1.466. These measures provide useful

insights into the performance of the model, enabling comparisons with competing models or forecasting methodologies.

In addition, the outcomes of the diagnostic Ljung-Box test, which was conducted on the residuals obtained from the ARIMA model utilizing the test dataset, demonstrate a significant X-squared value of 202.28 with 20 degrees of freedom. The p-value associated with the test is significantly small, considerably less than 0.05, suggesting substantial evidence to reject the null hypothesis that there is no serial correlation in the residuals. The discovery implies the possible existence of unexplained serial correlation, emphasizing the significance of improving the model or considering alternative models that can more accurately represent the underlying temporal patterns in the time series data.

To summarize, this extensive examination highlights the complex characteristics of time series modeling and prediction. The significance of a thorough methodology in data analysis and model selection is underscored, considering the intricate interrelationships and dependencies present in the data.

3.3 Generalized Linear Model

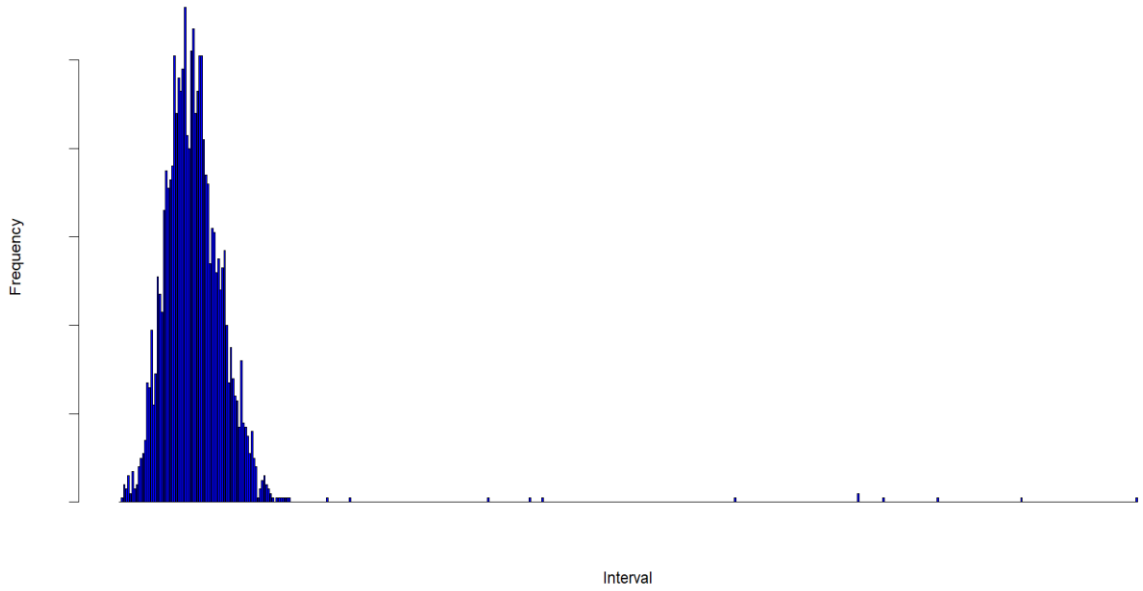
As the results weren't as expected in time series analysis it was made the decision to perform a GLM. The rationale behind transitioning from evaluating a time series dataset including 51 variables to employing a Generalized Linear Model (GLM) stemmed from the recognition that these numerous factors were not having a statistically significant influence on the observed temporal patterns and frequency within the dataset. The purpose of this strategy change is to analyze the influence of these 51 predictor variables on a specific response variable in a more targeted and understandable manner.

The application of the Generalized Linear Model (GLM) framework offers significant advantages, including improved interpretability and a more efficient modeling process. The implementation of this methodology allows for immediate evaluation of the relationships between the predictor variables and the response variable. The enormous size of the dataset presented a challenge in understanding the information during the initial time series analysis. The utilization of the Generalized Linear Model (GLM) facilitates the determination of statistically significant impacts of the 51 variables on the response variable. This specific skill helps the extraction of practical insights and assists in the process of making well-informed decisions. In summary, this transformation ultimately enhances the efficiency and accuracy of

the data exploration process, making the previously difficult dataset more appropriate for systematic analysis and focused research.

The initial analysis employed a methodology that involved examining a frequency distribution through the use of interval-based grouping. The aforementioned procedure resulted in the generation of a barplot, which functioned as a graphical depiction of the data's distribution. Based on a visual analysis of the barplot, it is apparent that the distribution displays a symmetrical, bell-shaped curve, without any conspicuous outliers or skewness. The visual characteristics seen strongly suggest that the data may adhere to a Normal distribution, displaying the typical characteristics associated with this statistical model. Nevertheless, it is essential to complement these visual observations with thorough statistical analyses to definitively establish the Normality of the dataset.

Figure 9 – Visual representation of frequency dataset



Then it was employed a multiple linear regression model to examine the relationship between the variable associated with the frequency and a group of predictor factors. The utilization of a Gaussian family for the regression model entails an inherent assumption of a normal distribution of data. This assumption is based on a thorough visual analysis of the data, which displayed features that are indicative of a normal distribution. These features include symmetry, the lack of prominent outliers, and the absence of any significant skewness.

Figure 10 – RStudio code that estimates the GLM

```
model <- glm(Dados$Frequencia ~ ., family = 'gaussian', data = Dados[, c("Frequencia", cols_to_impute)])
```

The results of the regression model, as presented in the output, provide significant insights into the relationships between the predictor variables and the response variable, frequency. The calculated coefficients for each predictor variable provide information about the magnitude and direction of their association with the response variable. The statistical significance of each coefficient is evaluated by examining the associated p-values, where smaller p-values suggest a stronger influence on the response variable.

In addition, the Akaike Information Criterion (AIC) value is provided as an indicator of the model's level of appropriateness. Smaller AIC values indicate a more optimal model fit to the given data. In the present case, the AIC value of 55897 indicates a satisfactory level of model fit to the data.

The final output is the number of iterations performed using the Fisher Scoring method for parameter estimate by maximum likelihood estimation.

In summary, the selection of the Gaussian distribution as the basis for the regression model, which is supported by the visual examination of data normality, enables the meaningful interpretation of coefficients, p-values, AIC, and the count of Fisher Scoring rounds. These results provide significant insights into the links among the data and the level of accuracy of the model in regard to the dataset, hence enhancing the overall comprehension of the examined data.

After presenting the generalized linear model (GLM) (check Appendix II and Appendix III for the output of the GLM), the focus now shifts to the examination of model coefficients and the statistical significance of variables. These aspects are essential components for comprehending the predictive capability of the model. The coefficients obtained from the Generalized Linear Model (GLM) provide crucial information regarding the impact of individual factors on the response variable, including both the direction and amount of their influence. Positive coefficients are indicative of positive relationships, whilst negative coefficients are indicative of negative relationships. Furthermore, the magnitude of the coefficient represents the intensity of the impact, where factors with bigger coefficients exert a more significant influence on the response variable. In addition to these coefficients, p-values play a crucial role in measuring the statistical significance of variables in predicting the

response. In general, p-values that are lower than the threshold of 0.05 are considered to be indicative of statistical significance. The purpose of this analysis is to provide a comprehensive understanding of the importance of each variable in explaining the fluctuations observed in the response variable. This will establish a solid foundation for conducting additional examination and evaluation of the model.

In the ongoing examination of the GLM output, a crucial step is undertaken to identify variables for inclusion based on their respective p-values. In this particular instance, variables that possess p-values below 0.1 are identified as having a significant influence. The aforementioned selection procedure serves to differentiate and prioritize the most significant variables within the model, so effectively mitigating extraneous influences by rejecting those that are less pertinent. The list of significant variables obtained provides a succinct overview of the primary predictors inherent in the model. The aforementioned selection process serves to improve the accuracy of the model, thereby facilitating a more thorough comprehension of the inherent relationships between variables and the response variable (check Appendix IV for the list of the variables with p-value higher then 0.1).

As research advances, it becomes evident that predictive modeling encompasses more than just the determination of variable significance. One crucial factor to take into account pertains to the distribution of model predictions. The aforementioned code utilizes the Generalized Linear Model (GLM) to produce predictions and subsequently stores these predictions for subsequent analysis. Evaluating the dispersion of these predictions is crucial for verifying the model's assumptions and overall dependability.

The existence of a set of predictions that follows a normal distribution suggests that the model's estimations are unbiased and exhibit good calibration. The assessment can be performed using a range of methodologies, encompassing statistical tests and visual representations such as histograms or quantile-quantile plots. The act of ensuring that the predictions made by the model conform to a normal distribution serves to bolster trust in its efficacy, hence strengthening its capacity to extract significant insights from the data and facilitate well-informed decision-making (check Appendix V for the histogram of the predictions made).

Chapter 4: Conclusions

This chapter provides a summary of the primary findings and implications of the conducted research. The research utilized a thorough technique that included data preprocessing, time series analysis, and the implementation of a Generalized Linear Model (GLM). The main aim of this study was to have a deeper understanding of the various elements that impact the frequency of Médis health insurance clients in Portugal.

The study commenced by conducting a comprehensive data gathering procedure from the Sales Index and PorData databases, which house crucial socio-economic indicators at the local level in Portugal. The dataset obtained consisted of 51 variables, which were methodically classified according to spatial and temporal criteria. The careful and thorough handling of the data established the foundation for further analysis.

The first step involved doing a time series analysis, where the ARIMA model was used to detect any temporal dependencies present in the dataset. While this particular methodology offered interesting insights into the temporal patterns under investigation, it is important to note that the conclusions obtained were not conclusive. This lack of conclusiveness can be attributed to the complexity present within the dataset.

A shift was made towards employing a Generalized Linear Model (GLM) methodology as a means of addressing the difficulties discovered throughout the time series analysis. This transition facilitated a more efficient analysis of the associations between predictor variables and the comsuption of Médis health insurance customers.

The GLM analysis revealed a selection of predictor variables that exhibited a substantial impact on customer frequency. The aforementioned findings provide valuable insights for healthcare insurers, such as Médis, which may be utilized to design customized tactics aimed at predicting customer frequency.

Moreover, the analysis of the histogram depicting the predictions generated by the Generalized Linear Model (GLM) revealed that the model's estimations adhered to a normal distribution. This implies that the model exhibits strong calibration and possesses the capacity to generate predictions that are without bias.

Nevertheless, it is crucial to recognize the constraints of the study. The dataset's unique intricacy posed difficulties in accurately capturing its subtleties using both time series analysis and GLM modeling techniques. The analysis conducted may not have adequately accounted for

certain subtle elements, highlighting the necessity for more investigation with more sophisticated modeling methodologies.

Furthermore, it should be noted that there were specific predictor variables that were not consistently available throughout different years and municipalities. This lack of data availability may have had an impact on the overall comprehensiveness of the research.

Looking forward to this research provides opportunities for future inquiries. Researchers may opt to utilize sophisticated modeling techniques, such as machine learning algorithms, in order to capture the deep relationships more fully, present within the dataset. The incorporation of supplementary data sources has the potential to augment the comprehension of exterior aspects that impact the frequency of private healthcare client utilization. By extending the study beyond the year 2022, a more comprehensive investigation of long-term trends and their implications might be conducted.

In summary, the used study approach has generated significant findings regarding the factors influencing the frequency of Médis health insurance clients in Portugal. The shift from utilizing time series analysis to employing the Generalized Linear Model (GLM) framework highlights the significance of flexibility and adaptability in the core of the research. The selection of salient variables and the establishment of a properly calibrated model has practical consequences for insurer professionals. Acknowledging the inherent limits of the study, it provides a robust basis for subsequent inquiries in this domain, presenting opportunities for sophisticated modeling and comprehensive data integration.

References

1. Acharya, A., Vellakkal, S., Taylor, F., Masset, E., Satija, A., Burke, M., & Ebrahim, S. (2013). The impact of health insurance schemes for the informal sector in low-and middle-income countries: a systematic review. *The World Bank Research Observer*, *28*(2), 236-266.

2. Agresti, A. (2007). An introduction to categorical data analysis. Hoboken, NJ: Wiley-interscience.

3. Beebe, A. (2021). *Impact of Urban Residuals-Based Amendments on Soil Health, Crop Yield, and Nutritional Quality*. University of Washington.

4. Białowolski, P., Węziak-Białowolska, D., & VanderWeele, T. J. (2019). The impact of savings and credit on health and health behaviours: an outcome-wide longitudinal approach. *International Journal of Public Health*, *64*, 573-584.

5. Box, G. E. P. and Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control, San Francisco, Holden-Day.

6. Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370.

7. Cascio, W. E. (2018). Wildland fire smoke and human health. *Science of the total environment*, *624*, 586-595.

8. Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine*, *6*(1), 21-26.

9. Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine*, *6*(1), 21-26.

10. Hansen, P. R., & Lunde, A. (2014). Estimating the persistence and the autocorrelation function of a time series that is measured with error. *Econometric Theory*, *30*(1), 60-93.

11. Hassani, H., & Yeganegi, M. R. (2019). Sum of squared ACF and the Ljung–Box statistics. *Physica A: Statistical Mechanics and its Applications*, *520*, 81-86.

12. Howard, G., Bartram, J., Water, S., & World Health Organization. (2003). Domestic water quantity, service level and health.

13. Moynihan, R., Sanders, S., Michaleff, Z. A., Scott, A. M., Clark, J., To, E. J., Jones, M., Kitchener, E., Fox, M., Johansson, M., Lang, E., Duggan, A., Scott, I., & Albarqouni, L. (2020, October 28). Pandemic impacts on healthcare utilisation: a systematic review. medRxiv (Cold Spring Harbor Laboratory); Cold Spring Harbor Laboratory.

14. Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199-203.

15. Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W., et al. (1996). Applied linear statistical models.

16. Salatin, P., & Bidari, M. (2014). Effects of Inflation on Women's Health in Selected Middle-Income Countries. *African Journal of Health Economics, 3*, 39-51.

17. Slutsky, E. (1937): The summation of random causes as the source of cyclic processes, Econometrica, 5, 105-146.

18. Walker, G. T. (1931): On periodicity in series of related terms, Proceedings of Royal Society, London Ser. A, 131, 518-532.

19. Wold, H. (1938): A Study in the Analysis of Stationary Time Series, Uppsala, Sweden, Almquist and Wicksell.

20. Yule, G. U. (1921): On the time-correlation problem, with especial reference to the variate-difference correlation method, with discussion, Journal of the Royal Statistical Society, 84, 497-537.

21. Yule, G. U. (1926): Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series, Journal of the Royal Statistical Society, 89, 1-64.

22. Yule, G. U. (1927): On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, Philosophical Transactions of the Royal Society, A, 226, 267-226.

Appendix

Appendix I

Variable	Description	Source
AnoAnalise	Year referring to the policy	Data Base Médis
MesAnalise	Month referring to the policy	Data Base Médis
Desc_Tipo_Negocio_Tecnico	Description of the type of business	Data Base Médis
Count_of_Num_Medis	Number of people with policies in a certain year and month of analysis	Data Base Médis
Sum_of_Exposicao_PS_AnoCivil	Sum of exposure per insured person in a calendar year	Data Base Médis
Sum_of_Num_Claims_Pagavel	Sum of the number of payable claims	Data Base Médis
Concelho	Municipality regarding the policy	Data Base Médis
Valor_Med_Mensal	Number of doctors per inhabitant monthly	Sales Index
CS_Medio_Mensal	Number of health centers per municipality	Sales Index
SM_Medio_Mensal	Average salary by municipality	Sales Index
AC_Anual	The annual percentage of drinkable water in the municipality	Sales Index
Incendio_Ano	Percentage of area burned annually in a given municipality	Sales Index
Farm_Ano	Number of pharmacies per municipality	Sales Index
Dias uteis	Number of working days per month	Sales Index
Ind_Consumo	Annual consumption index by municipality	Sales Index
Ind_Pot_Bancos	Potential annual bank index by municipality	Sales Index
Ind_Pot_Farmacias	Annual potential pharmacy index by municipality	Sales Index
Ind_Aloj_Tur	Potential annual tourist accommodation index by municipality	Sales Index
Ind_Emp	Annual business index by municipality	Sales Index
Ind_Pot_Com	Potential annual trade index by municipality	Sales Index
Ind_Dist_Pop	Annual population distribution index by municipality	Sales Index
Ind_Geral_Saude	General annual health index by municipality	Sales Index
Ind_Ut_SegSocial	Index of annual social security users by municipality	Sales Index
Ind_Pensionistas	Annual pensioner index by municipality	Sales Index
Temperatura	Average annual temperature for each municipality	Sales Index
Res_Urb	Average annual urban waste per municipality and per inhabitant	Sales Index

Ind_Dep_Idosos	Elderly Dependency Index	Sales Index
Ind_Dep_Jovens	Youth Addiction Index	Sales Index
Ind_Envelhecimento	Aging Index	Sales Index
Ind_Longevidade	Longevity Index	Sales Index
Ind_Mulh_Fertil	Women of childbearing age	Sales Index
Pop_Ativa	Active Population	Sales Index
Tax_Brut_Mort	Crude Mortality Rate	Sales Index
Tax_Brut_Nat	Gross Birth Rate	Sales Index
Tax_Fec	Fertility Rate	Sales Index
Tax_Cres	Migration Growth Rate	Sales Index
Pop_Estr_Est_Res	Foreign Population with Legal Resident Status	Sales Index
Lares	Total number of homes per municipality	Sales Index
Centros_Dia	Total amount of Day Centers per municipality	Sales Index
Creche_Jardinfancia	Total number of Nurseries and Kindergartens per municipality	Sales Index
Pessoa_Por_Lar	Total number of people in homes per municipality	Sales Index
Pessoa_Por_Centro_Dia	Total number of people in Day Centers per municipality	Sales Index
Inflacao	Annual inflation	PorData
Inflacao_Saude	Annual health inflation	PorData
Desp_Benef_ADSE	Expense per beneficiary by ADSE	PorData
Desp_Per_Cap_SNS	Per capita expenditure related to the SNS	PorData
Camas_Hosp_Geral	Total number of beds in general hospitals	PorData
Camas_Hosp_Esp	Total number of beds in specialized hospitals	PorData
Condenados_Por_1000Hab	Number of convicts per 1000 inhabitants	PorData
Ind_Global_Bem_Estar	Global Wellbeing Index	PorData
Ind_Cond_Mat_Vida	Index of material living conditions	PorData
Ind_Qual_Vida	Quality of life index	PorData
Tax_Poupanca_Familia	Family savings rate	PorData
Tax_Cresc_Pib	GDP growth rate	PorData
Ferias	Months with school holidays	PorData
Lockdown	Months of lockdown due to the pandemic	PorData
Receitas medicas	Number of medical prescriptions assigned annually	PorData
Embalagens	Number of packages allocated annually	PorData

Appendix II

```
Call:
glm(formula = Dados$Frequencia ~ ., family = "gaussian", data = Dados[,
    c("Frequencia", cols_to_impute)])

Coefficients: (9 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -9.513e+00  2.307e+00  -4.124 3.74e-05 ***
VALOR_MED_MENSAL        2.921e-05  4.514e-05   0.647  0.51758
CS_ANUAL               -2.764e-02  1.227e-02  -2.253  0.02428 *
SM_MEDIO_MENSAL        -7.154e-05  4.087e-05  -1.751  0.08001 .
AC_ANUAL                8.124e+00  9.800e-01   8.290  < 2e-16 ***
INCENDIO_ANO            1.675e+01  4.010e+00   4.179 2.95e-05 ***
FARM_ANO                6.275e-03  4.178e-03   1.502  0.13313
Dias.uteis              1.018e-01  1.040e-02   9.784  < 2e-16 ***
IND_CONSUMO            -2.136e-03  1.064e-03  -2.008  0.04470 *
IND_POT_BANCOS          1.819e-03  4.268e-04   4.263 2.03e-05 ***
IND_POT_FARMACIAS      -3.564e-03  3.356e-04 -10.620  < 2e-16 ***
IND_ALOJ_TUR            2.051e-02  3.631e-03   5.648 1.65e-08 ***
IND_EMP                -3.638e-03  6.837e-04  -5.321 1.05e-07 ***
IND_POT_COM            -3.246e-03  6.904e-04  -4.701 2.61e-06 ***
IND_DIST_POP            1.019e-01  1.780e-02   5.726 1.05e-08 ***
IND_GERAL_SAUDE        -8.202e-02  1.017e-02  -8.063 7.96e-16 ***
IND_UT_SEGSOCIAL        4.069e-02  8.292e-03   4.908 9.30e-07 ***
IND_PENSIONISTAS        1.652e-02  2.502e-02   0.660  0.50908
IND_DEP_IDOSOS          1.386e-02  6.458e-03   2.146  0.03188 *
IND_DEP_JOVENS          1.056e-01  1.105e-02   9.553  < 2e-16 ***
IND_ENVELHECIMENTO     -2.213e-03  5.612e-04  -3.943 8.08e-05 ***
IND_LONGEVIDADE         8.550e-03  3.854e-03   2.219  0.02653 *
IND_MULH_FERTIL        -3.504e-02  1.284e-02  -2.728  0.00637 **
POP_ATIVA               6.791e-02  8.741e-03   7.769 8.42e-15 ***
TAX_BRUT_MORT          -2.240e-03  4.838e-03  -0.463  0.64340
TAX_BRUT_NAT           -2.388e-01  6.090e-02  -3.920 8.88e-05 ***
TAX_FEC                 3.774e-02  1.253e-02   3.011  0.00261 **
TAX_CRES                2.767e-01  2.490e-02  11.114  < 2e-16 ***
POP_ESTR_EST_RES       -2.417e-05  4.994e-06  -4.839 1.32e-06 ***
LARES                  -9.243e-04  3.311e-03  -0.279  0.78012
CENTROS_DIA            -4.842e-02  3.841e-03 -12.607  < 2e-16 ***
CRECHE_JARDINFANCIA    -3.891e-03  2.377e-03  -1.637  0.10163
PESSOA_POR_LAR         -4.551e-03  7.844e-04  -5.802 6.67e-09 ***
PESSOA_POR_CENTRO_DIA   1.399e-02  2.013e-03   6.947 3.89e-12 ***
INFLACAO                1.841e-01  2.220e-02   8.293  < 2e-16 ***
INFLACAO_SAUDE          2.625e-01  4.519e-02   5.808 6.44e-09 ***
DESP_BENEF_ADSE         6.565e-03  7.929e-04   8.280  < 2e-16 ***
DESP_PER_CAP_SNS        3.275e-04  2.978e-04   1.100  0.27138
CAMAS_HOSP_GERAL       -2.863e-04  1.089e-04  -2.628  0.00859 **
CAMAS_HOSP_ESP                 NA         NA      NA       NA
CONDENADOS_POR_1000HAB         NA         NA      NA       NA
IND_GLOBAL_BEM_ESTAR           NA         NA      NA       NA
IND_COND_MAT_VIDA              NA         NA      NA       NA
IND_QUAL_VIDA                  NA         NA      NA       NA
TAX_POUPANCA_FAMILIA           NA         NA      NA       NA
TAX_CRESC_PIB                  NA         NA      NA       NA
Ferias                 -3.175e-01  2.420e-02 -13.120  < 2e-16 ***
Lockdown               -1.704e+00  7.772e-02 -21.926  < 2e-16 ***
Receitas.medicas               NA         NA      NA       NA
Embalagens                     NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.108499)
```

Appendix III

```
    Null deviance: 44611  on 15584  degrees of freedom
Residual deviance: 32775  on 15544  degrees of freedom
AIC: 55897

Number of Fisher Scoring iterations: 2
```

Appendix IV

	Variable	Coefficient	P_Value
(Intercept)	(Intercept)	-9.513210e+00	3.741472e-05
CS_ANUAL	CS_ANUAL	-2.764261e-02	2.428430e-02
SM_MEDIO_MENSAL	SM_MEDIO_MENSAL	-7.154515e-05	8.001230e-02
AC_ANUAL	AC_ANUAL	8.123846e+00	1.224653e-16
INCENDIO_ANO	INCENDIO_ANO	1.675435e+01	2.949498e-05
Dias.uteis	Dias.uteis	1.018001e-01	1.529962e-22
IND_CONSUMO	IND_CONSUMO	-2.136066e-03	4.469866e-02
IND_POT_BANCOS	IND_POT_BANCOS	1.819398e-03	2.031689e-05
IND_POT_FARMACIAS	IND_POT_FARMACIAS	-3.564145e-03	2.962570e-26
IND_ALOJ_TUR	IND_ALOJ_TUR	2.050931e-02	1.650334e-08
IND_EMP	IND_EMP	-3.637582e-03	1.047346e-07
IND_POT_COM	IND_POT_COM	-3.245634e-03	2.607630e-06
IND_DIST_POP	IND_DIST_POP	1.019172e-01	1.046513e-08
IND_GERAL_SAUDE	IND_GERAL_SAUDE	-8.201805e-02	7.955061e-16
IND_UT_SEGSOCIAL	IND_UT_SEGSOCIAL	4.069466e-02	9.304218e-07
IND_DEP_IDOSOS	IND_DEP_IDOSOS	1.386010e-02	3.188232e-02
IND_DEP_JOVENS	IND_DEP_JOVENS	1.055894e-01	1.439265e-21
IND_ENVELHECIMENTO	IND_ENVELHECIMENTO	-2.212763e-03	8.080371e-05
IND_LONGEVIDADE	IND_LONGEVIDADE	8.550358e-03	2.652887e-02
IND_MULH_FERTIL	IND_MULH_FERTIL	-3.503535e-02	6.373601e-03
POP_ATIVA	POP_ATIVA	6.790506e-02	8.416266e-15
TAX_BRUT_NAT	TAX_BRUT_NAT	-2.387606e-01	8.877066e-05
TAX_FEC	TAX_FEC	3.773819e-02	2.605943e-03
TAX_CRES	TAX_CRES	2.766974e-01	1.368657e-28
POP_ESTR_EST_RES	POP_ESTR_EST_RES	-2.416801e-05	1.316797e-06
CENTROS_DIA	CENTROS_DIA	-4.842122e-02	2.911084e-36
PESSOA_POR_LAR	PESSOA_POR_LAR	-4.551398e-03	6.671907e-09
PESSOA_POR_CENTRO_DIA	PESSOA_POR_CENTRO_DIA	1.398701e-02	3.889380e-12
INFLACAO	INFLACAO	1.840897e-01	1.189428e-16
INFLACAO_SAUDE	INFLACAO_SAUDE	2.624943e-01	6.439735e-09
DESP_BENEF_ADSE	DESP_BENEF_ADSE	6.565250e-03	1.332742e-16
CAMAS_HOSP_GERAL	CAMAS_HOSP_GERAL	-2.862699e-04	8.588157e-03
Ferias	Ferias	-3.175432e-01	4.060794e-39
Lockdown	Lockdown	-1.704120e+00	5.676143e-105

Appendix V