UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Self-reported Diabetes in Portugal and its association with health-related quality of life and medical resources consumption using a nationwide prospective cohort

Vanessa Sofia Salgueiro Lopes

Mestrado em Bioestatística

Trabalho de projeto orientado por:
Professora Doutora Marília Antunes
Professora Doutora Ana Maria Rodrigues

2022

"All models are wrong, but some are useful"

George E. P. Box

"The only way to discover the limits of the possible is to go beyond them into the impossible"

Arthur C. Clarke

"All our dreams can come true, if we have the courage to pursue them"

Walt Disney

Agradecimentos

Com muito esforço e dedicação elaborei este projeto, mas nunca o conseguiria fazer sozinha. Quero agradecer aos meus pais, Carla e Fernando, por tudo aquilo que me deram para que pudesse chegar até aqui. Por todo o apoio que demonstraram e que compreenderam sempre quando tinha de substituir os momentos familiares pela faculdade. Agradecer também ao meu irmão Rodrigo, porque apesar de não compreender tão bem esta dedicação ao lado escolar, sempre esteve presente para mim. Aos meus pais e ao meu irmão, agradeço a paciência que tiveram nos momentos de stress que um projeto destes implica.

Um agrecimento especial para as minhas orientadoras, Professora Marília Antunes e Professora Ana Rodrigues. À professora Marília agradeço toda a paciência, toda a ajuda nos momentos de aflição. Tenho uma admiração pela capacidade de trabalho da professora Marília e pela sua sabedoria, foi muito gratificante fazer este projeto com a sua colaboração. À professora Ana Rodrigues, quero agradecer por me ter permitido trabalhar neste projeto, por me ter integrado tão bem na equipa do CEDOC. Agradeço por ter sido sempre muito acessível comigo e por toda a disponibilidade.

Agradecer também a todos os professores do departamento de Estatística da Faculdade de Ciências da Universidade de Lisboa que se cruzaram no meu percurso e me ajudaram sempre em tudo o que precisei.

A todos os elementos da equipa do CEDOC, com um agradecimento especial à Rita, por ter estado sempre disponível para me ajudar.

À minha amiga Jéssica, agradeço todo o companheirismo, todas as horas de estudo, de aflição, de encorajamento, de partilha de conhecimento, de amizade. O meu percurso universitário não seria o mesmo sem a presença dela e sem as suas palavras de conforto. Como ela diz "O que a faculdade uniu, ninguém separa".

À minha amiga Leonor, por me ter ajudado nos momentos em que mais precisava, por nunca ter deixado que desistisse. Por ter feito parte deste percurso como ninguém, por saber que apesar de segunda escolha, sempre fui a primeira. Obrigada por cada momento.

Agradeço a todos os meus amigos e família que estiveram presentes neste percurso que vai deixar saudades.

Resumo

A diabetes é uma doença metabólica crónica caracterizada por níveis elevados de glucose no sangue. A quantidade de glucose é também chamada glicose e deve estar presente no sangue entre 70 a 110 miligramas por decilitro num indivíduo saudável. A glucose é a principal fonte de energia do organismo e resulta da digestão e transformação de alimentos como os amidos e açúcares presentes na dieta. A insulina é necessária para que possa ser utilizada como fonte de energia. A insulina, conhecida como a hormona da vida, é produzida nas células do pâncreas e é segregada por ela. É responsável pelo controlo dos níveis de glicose no sangue (Brutsaert, 2020).

Os principais tipos de diabetes são a diabetes tipo 1 e a diabetes tipo 2. A diabetes tipo 1 é também conhecida como diabetes juvenil (é mais comum em crianças e jovens) ou diabetes insulino-dependente. Acontece quando o pâncreas produz pouca ou nenhuma insulina por si só. Assim, a glicose permanece no sangue, fazendo com que os níveis de glicose aumentem, o que pode levar à hiperglicemia. A diabetes de tipo 2 é a forma mais comum da doença e ocorre normalmente em adultos. Acontece quando o corpo se torna resistente à insulina ou produz insulina em quantidade insuficiente (Lusiadas, 2022).

De acordo com o relatório do Observatório Nacional de Diabetes, em 2018 a prevalência da diabetes era de 13.6% em indivíduos com idades compreendidas entre os 20 e 79 anos (7 milhões e 700 mil pessoas) (Raposo, 2020). Portanto, estima-se que o número de diabéticos tenha sido de aproximadamente 1 milhão e 50 mil entre estas idades. A prevalência da diabetes tem vindo a aumentar ao longo dos anos. É um grande fardo para o sistema nacional de saúde, uma vez que traz problemas graves a ele associados ao longo do tempo, tais como problemas cardíacos, oculares, renais, nervosos e dos vasos sanguíneos, podendo aumentar o consumo de recursos de saúde. A diabetes perturba muito a vida dos indivíduos estando associada a pior qualidade de vida e ao aumento da mortalidade precoce. Por conseguinte, é importante caracterizar a população portuguesa de diabéticos bem como as comorbilidades que mais frequentemente acompanham a diabetes. (Organization, 2000).

Este projeto visa caracterizar os diabéticos adultos portugueses e o impacto da diabetes na qualidade de vida e nos recursos de saúde. Este estudo faz parte de um projeto proposto pela Unidade EpiDoC da Nova Medical School, onde foi desenvolvido em 2011 um estudo de coorte prospetivo intitulado Epidemiologia em Doenças Crónicas (EpiDoC). O estudo EpiDoC foi concebido por investigadores da Faculdade de Medicina NOVA, em Lisboa, com uma estreita colaboração entre o social e cientistas biomédicos, assegurando uma abordagem multidisciplinar completa. Para este fim, o estudo foi concebido para ser representativo da população adulta portuguesa. O EpiDoC é um estudo prospetivo de coorte fechado, incluindo uma amostra nacional representativa dos adultos (maiores de 18 anos) que não eram institucionalizados e viviam em casas particulares em Portugal Continental e Ilhas (Açores e Madeira).

O principal objetivo deste projeto foi caracterizar a população dibética em adultos portugueses uti-

lizando a coorte EpiDoC. Os dados recolhidos nas três ondas de seguimento foram analisados para avaliar o impacto da diabetes na perda de qualidade de vida e no consumo de recursos médicos, nomeadamente na ocorrência de hospitalizações e no número de consultas médicas. Os objetivos específicos foram os seguintes:

- Caracterizar a população relativamente à diabetes auto-reportada entre os adultos portugueses, tendo em conta a distribuição geográfica, características sócio-demográficas e comorbilidades associadas à doença. Para isso, foi efetuada uma análise exploratória alargada, na qual se recorreu a representações gráficas, ao cálculo de prevalências, ao cálculo de medidas de localização e dispersão e à utilização de regras de associação e de agrupamento.

- Avaliar o impacto da diabetes na qualidade de vida - medido pelo score EuroQol Five Dimensional Questionnaire 3 Level Version (EQ-5D-3L) - em cada onda de recolha de dados, bem como numa perspetiva longitudinal, a fim de avaliar a evolução desta medida ao longo do tempo. Para responder a esta questão foi utilizado o modelo de regressão Tobit.

- Avaliar o impacto da diabetes na ocorrência de internamentos hospitalares e no número de consultas médicas através de modelos lineares generalizados longitudinais para dados binários e contagens, respetivamente. Para estas últimas análises foram usados os modelos GEE com os respetivos ajustes ao modelo para se adequar ao tipo de variável resposta de cada abordagem.

Com recurso às prevalências, e analisando as respostas dos indivíduos na primeira onda, conclui-se que a proporção de diabéticos é superior no sexo feminino comparando com o sexo masculino, ainda assim, a diferença entre diabéticos e não diabéticos quanto ao sexo não é significativa. Para além disto, existe uma maior prevalência de diabéticos de etnia caucasiana, na faixa etária dos 66 aos 75 anos e casados. Relativamente às NUTS II, existe uma maior proporção de diabéticos na região Norte. Quanto ao nível de educação, observa-se uma maior proporção de diabéticos cuja escolaridade varia entre 1 e 4 anos. No entanto, a idade pode ser considerada um fator de confundimento, uma vez que existe uma maior probabilidade de baixos níveis de educação estarem associados a indivíduos com uma idade mais avançada. O facto da maioria dos diabéticos estarem reformados, corrobora com a ideia de a idade ser um fator de confundimento. No que diz respeito aos hábitos de vida, conclui-se que a maioria dos diabéticos ingerem álcool, não fumam e praticam exercício. Tal como esperado, a maioria dos diabéticos tem um Índice de Massa Corporal (IMC) correspondente à categoria de obesidade. Por fim, a pressão arterial elevada mostrou ser a doença mais prevalente nos diabéticos.

Com recurso a modelos da classe dos modelos lineares generalizados, respondeu-se às questões sobre o impacto da diabetes na qualidade de vida. Analisaram-se as estimativas dos parâmetros associados às variáveis, com o objetivo de averiguar se estavam de acordo com o mencionado na literatura. Os indivíduos foram seguidos ao longo do tempo e concluí-se que a qualidade de vida diminui com a presença da diabetes. Esta conclusão corrobora com outros artigos publicados, uma vez que a qualidade de vida piora com a presença da diabetes, pelo que esta doença acarreta (Brown et al., 2000).

Na ocorrência de internamentos hospitalares, numa primeira análise averiguou-se que nos indivíduos diabéticos apenas 14.5% tinham sido hospitalizados. Através da análise do modelo GEE com abordagem longitudinal concluí-se que, ao longo do tempo, a chance de um indivíduo ser hospitalizado aumenta 30% nos indivíduos com diabetes, comparando com os não diabéticos. Esta conclusão está de acordo com o mencionado no Relatório anual do observatório nacional da Diabetes, que refere "o número de utentes saídos/internamentos em que a Diabetes surge como diagnóstico associado tem evidenciado uma

dinâmica de crescimento acentuada...” (Diabetes, 2016).

Relativamente ao modelo referente ao número de consultas, este apresentou alguns problemas com os resíduos, uma vez que estes não se apresentavam de uma forma muito regular. Pela análise das variáveis explicativas contra o número de consultas, pode constatar-se que as categorias de cada variável diferem muito pouco no que diz respeito ao número de consultas. Ainda assim, analisou-se o modelo para se perceber se ia ao encontro do mencionado na literatura, o que se verificou. A presença da diabetes faz com que o número esperado de consultas aumente.

Palavras-chave: Diabetes; Modelos GEE; Modelos GLM; Modelos Tobit; Prevalência.

Abstract

Diabetes is a chronic metabolic disease characterised by high blood glucose levels. It is a major burden on the national healthcare system as it brings serious problems associated with it over time, such as heart, eye, kidney, nerve and blood vessel problems, and can increase the consumption of healthcare resources. Diabetes greatly disrupts the lives of individuals being associated with poorer quality of life and increased early mortality. Therefore, it is important to characterise the Portuguese diabetic population as well as the comorbidities that most often accompany diabetes. (Organization, 2000).

This study is part of a project proposed by EpiDoC Unit at Nova Medical School, where a prospective cohort study entitled Epidemiology in chronic diseases (EpiDoC) was developed in 2011. To this end, this study was designed to be representative of the Portuguese adult population. EpiDoC is a prospective closed cohort study, including a nationally representative sample of adults (older than 18 years) who were not institutionalised and lived in private homes in Mainland Portugal and Islands (Açores and Madeira).

The main objective of this project was to characterise the diabetic population in Portuguese adults using the EpiDoC cohort. Data collected at the three follow-up waves were analysed to assess the impact of diabetes on the loss of quality of life and on the consumption of medical resources, namely the occurrence of hospitalisations and the number of medical appointments. The specific objectives were as follows:

- To characterise the population regarding self-reported diabetes among Portuguese adults, taking into account the geographical distribution, socio-demographic characteristics and comorbidities associated with the disease. To this end, a broad exploratory analysis was conducted.
- To assess the impact of diabetes on quality of life - measured by the EuroQol Five Dimensional Questionnaire score - at each data collection wave, as well as in a longitudinal perspective, in order to assess the evolution of this measure over time. To answer this question, the Tobit regression model was used.
- To assess the impact of diabetes on the occurrence of hospital admissions and the number of medical appointments through longitudinal generalized linear models for binary data and counts, respectively.

By the analysis of the prevalences, and analysing the responses of individuals in the first wave, it is concluded that there are more diabetic women than men, yet this difference is not significant. Furthermore, there is a higher prevalence of diabetics of Caucasian ethnicity, aged 66 to 75 years and married. Regarding NUTS II, there is a higher proportion of diabetics in the Norte region. Regarding the level of education, there is a higher proportion of diabetics whose schooling varies between 1 and 4 years. However, age may be considered a confounding factor, since there is a higher probability of low levels of education being associated with individuals with a more advanced age. The fact that most diabetics

were retired corroborates the idea of age being a confounding factor. With regard to lifestyle habits, we concluded that most diabetics drink alcohol, do not smoke and do exercise. As expected, most diabetics have a Body Mass Index (BMI) corresponding to the obesity category. Finally, high blood pressure proved to be the most prevalent disease in diabetics.

Regarding the quality of life models, it decreases with the presence of diabetes. As regards the occurrence of hospital admissions, a first analysis showed that only 14.5% of diabetic individuals had been hospitalized. Through the analysis of the model, concluded that the chance of an individual being hospitalized increases in diabetic patients. With regard to the number of consultations, it was found that diabetes causes the expected number of medical appointments to increase.

Keywords: Diabetes; GEE Models; GLM Models; Tobit Models; Prevalence.

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

ANOVA Analysis of Variance
BMI Body Mass Index
CEDOC Centro de Estudos de Doenças Crónicas (Chronic Diseases Research Center)
EpiDoC Epidemiologia em doenças crónicas
EQ-5D EuroQol Five Dimensional Questionnaire
GEE Generalized estimating equation
GLM Generalized linear model
GLME Generalized Linear Mixed-Effects Models
NUTS Nomenclature of territorial units for statistics
PSU Primary Sampling Unit Randomization
RMD Rheumatic and Musculoskeletal Diseases
SNS Serviço Nacional de Saúde
OR Odds Ratio

1. Introduction

Diabetes is a chronic metabolic disease characterised by high blood glucose levels. In Portugal, according to the National Diabetes Observatory report, in 2018 the prevalence of diabetes was 13.6% in individuals aged between 20 and 79 years (7700000 people) (Raposo, 2020). In addition, diabetes is a major burden on the Serviço Nacional de Saúde (SNS) as it brings serious problems associated with it over time, such as heart, eye, kidney, nerve and blood vessel problems.

One of the aims of this study is to characterise the population regarding self-reported diabetes among Portuguese adults, taking into account the geographical distribution and socio-demographic characteristics. To this end, an extended exploratory analysis was carried out using graphical representations, the calculation of measures of location and dispersion and the calculation of weighted prevalence. Another objective of this project was to assess the impact of diabetes on quality of life - measured by the EuroQol Five Dimensional Questionnaire score - at each data collection wave, as well as in a longitudinal perspective, in order to assess the evolution of this measure over time. Since the variable quality of life score has a mass point at 1, this variable is right censored, as values greater than 1 are not allowed. To handle these types of variables, models that admit censoring are used and for this issue the Tobit regression model was used. Finally, the last objectives were to assess the impact of diabetes on the occurrence of hospital admissions and on the number of medical appointments through longitudinal generalized linear models for binary data and counts, respectively. Through GLM two approaches could be used: GLME or GEE. The GEE models were the models applied.

The variables used in the course of the project were collected through the EpiDoC study, with the purpose of answering several questions. It should be noted that the purpose of this data collection was not based on the specific analysis of diabetes. Still, as diabetes is a chronic disease that entails several other health problems, it is important to study this disease, which has been increasing over the years.

The variables used were as follows:

- Sociodemographic variables: sex, age group, NUTS II, ethnicity, marital status, education leval, employment status.
- Lifestyle related variables: alcohol intake, smokig habit, exercise practice.
- Health related variables: diabetes, cholesterol, cardiac, high blood pressure, mental, allergies, rheumatic, gastrointestinal, pulmonary, BMI.
- Quality of life variables: score EQ-5D.
- Medical resource consumption variables: hospitalizations, number of medical appointments.
- Other variables: wave, time of exposure.

All analyses were developed in R software, except the quality of life score model for the longitudinal approach, as there were limitations in R software for the development of this model. Stata software was used as an alternative.

This report is organized as follows: Chapter 2 contains the methodologies needed to answer the research questions. Chapter 3 describes data collection and exploratory data analysis. Chapter 4 analyses the impact of diabetes on quality of life through the questionnaire of quality of life, on the occurrence of hospital admissions and on the number of medical appointments. Finally, chapter 5 presents the conclusions.

2. Methodology

In order to group diseases, an association rule technique and an analysis through distance matrices were used. The variables of interest in this project are of different natures. One of the variables is quantitative and has an accumulation point. Another is a binary variable and finally a count variable. Hence, different types of approaches will be used, namely Tobit, Logistic and Poisson regression models.

2.1 Association rules technique

Association rules are a rule-based machine learning method for finding associations and relationships of interest between variables in large datasets. There are several metrics for understanding the strength of association between sets (Garg, 2018).

- Support: This measure gives an idea of the frequency of a set of items in all combinations. If a set of combinations has very little support, it is not possible to have sufficient information about the relationship between the variables and therefore no conclusions can be drawn from such a rule.
  $\text{Support}(\{X\} \rightarrow \{Y\}) = \frac{\text{Number of combinations that contain X and Y}}{\text{Total number of combinations}}$.
- Confidence: This measure defines the probability of occurrence of a given event knowing that other have already occurred.
  $\text{Confidence}(\{X\} \rightarrow \{Y\}) = \frac{\text{Number of combinations that contain X and Y}}{\text{Total number of combinations that contain X}}$.
- Coverage: This measure represents the frequency of the previous set happens. It is basically a support of the predecessor.
- Count: Absolute frequency of each set.

2.2 Distance matrix and dendrogram

A distance matrix contains the distances between pairs of objects. In data analysis, distance matrices are mainly used when hierarchical clustering is intended (Bock, n.d.).

Distance matrices are calculated using raw data. Given a raw data matrix:

Table 2.1: Raw data.

	X	Y
A	a	b
B	c	d

Through the raw data Euclidean distances are calculated. By Table 2.1 the distance between objects A and B would be calculated through characteristics X and Y, as follows:

$$\sqrt{(c-a)^2+(d-b)^2} = e \tag{2.1}$$

This would produce the following distance matrix:

Table 2.2: Distance matrix.

	A	B
A	0	e
B	e	0

The distance matrix can be reproduced graphically, the most common case being the use of the dendrogram. There are several methods to elaborate the dendrograms, in this case only the Ward method is presented.

Ward's method consists in a grouping of the data forming groups in order to always reach the smallest internal error between the vectors, which compose each group, and the centroid of the group. This method is based on the calculation of the average dissimilarity, given by the expression (Mendes Leal, 2019):

$$\frac{n_k n_{k'} d^2(\bar{x}_k, \bar{x}_{k'})}{n_k + n_{k'}} \tag{2.2}$$

This measure of average distance is equivalent to the increment that the sum of the squares of the distances of the elements of the classes from their centroids, undergoes when two classes C_K and $C_{K'}$ are joined. The average dissimilarity causes these distances to increase.

$$I_{C_K C_{K'}} = \frac{n_k n_{k'}}{n_k + n_{k'}} \sum_{\ell=1}^{p} (\bar{x}_{\ell k}, \bar{x}_{\ell k'})^2 \Leftrightarrow I_{C_K C_{K'}} = \frac{n_k n_{k'}}{n_k + n_{k'}} d^2(\bar{x}_k, \bar{x}_{k'}) \tag{2.3}$$

2.3 Tobit Regression Models

In scenarios where the response variable is continuous quantitative but strictly constrained between two values, linear regression models are not suitable. Beta regression models could be applied, however, when there is an accumulation point at the boundaries of the interval, these models are also not suitable. Therefore, the solution is to use Tobit models (Smith and Brame, 2003, Klein and Moeschberger, 1997).

James Tobin proposed a new model in a study published in 1958. This model is a linear regression model in which the relationships of a bounded dependent variable are estimated. The observations of the response variable are incomplete due to some type of censoring, that is, they are not represented with their real value but with the censored value.

The Tobit regression model is comparable to the classical linear regression model:

$$Y_i^* = \mathbf{X_i}' \beta + \varepsilon_i, \; i = 1, ..., n, \tag{2.4}$$

where,

- Y^*: Latent dependent variable.

- $\mathbf{X}_i = [X_{1i}, X_{2i}, X_{3i}, ..., X_{ki}]$
  Observed vector of explanatory variables.

- $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}$
  Regression coefficients associated with the explanatory variables.

- ε_i: independent and identically distributed random variables.
  $\varepsilon \sim N(0, \sigma^2)$

Tobit regression applies in two situations, when truncation occurs and when censoring occurs.

2.3.1 Truncation

Truncation exists when, due to a selection process inherent to the study design, only individuals to whom a certain event occurs are studied. With this, some observations become unavailable in the response variable and in the explanatory variables.

2.3.2 Censoring

Censoring occurs when for certain observations of the sample the data on the response variable are not available in their entirety, because they are limited, but unlike truncation, data on the explanatory variables is available. Censoring happens when it is not possible to measure the observations due to some kind of limitation. Then the values contained in this limitation area are all reported with the same value.

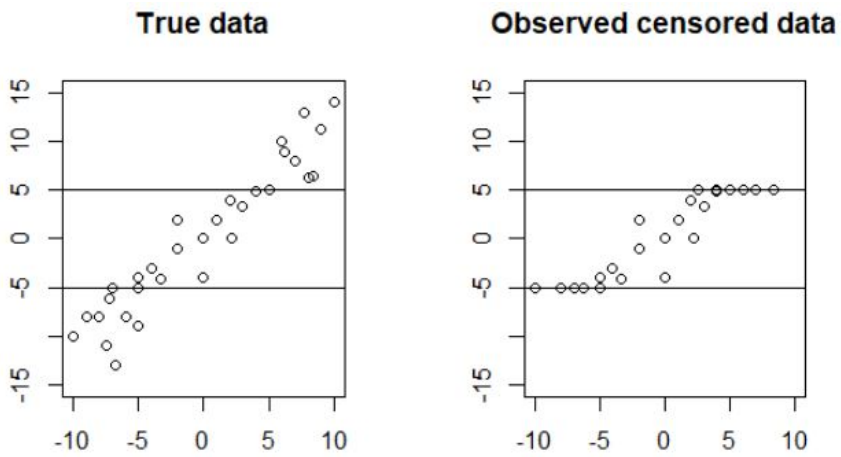

Fig. 2.1: Example of censored data.

In Figure 2.1 it can be seen that the data is limited between -5 and 5. These are also the mass points. Therefore, all values below -5 and above 5 take these values, respectively. As shown in the figure on the right.

In the censored regression the following is observed (censorship on the left and right, respectively):

$$y_i = \begin{cases} c, & y_i^* \leqslant c \\ y_i^*, & y_i^* > c \end{cases} \qquad y_i = \begin{cases} c, & y_i^* \geqslant c \\ y_i^*, & y_i^* < c \end{cases}, \tag{2.5}$$

where,
y_i is the observed variable,
$yi*$ is the latent variable, that is, the true value of each observation, and
c is the accumulation point.

Considering the right censored case, substituting (2.4) in (2.5),

$$y_i = \begin{cases} c, \ \ if \ \ y^* = \sum_{j=0}^{p} \beta_j x_{ij} + \varepsilon_i \geqslant c \\ y^*, \ \ if \ \ y^* = \sum_{j=0}^{p} \beta_j x_{ij} + \varepsilon_i < c \end{cases} \tag{2.6}$$

The probability density function of the censored variable is given by:

$$g(y) = \begin{cases} f(y^*), \ \ if \ \ y < c \\ 1 - F(c), \ \ if \ \ y = c \end{cases} \tag{2.7}$$

Being,

- $f(y^*)$ the latent variable's pdf
- $1 - F(y^*) = S(y)$ the survival function
- $d_i = \begin{cases} 1, y_i < c \\ 0, y_i = c \end{cases}$

the likelihood function for the censored variable is given by

$$\mathscr{L}(y,d) = \prod_{i=1}^{n} f(y_i)^{d_i} S_f(y_i)^{1-d_i}. \tag{2.8}$$

Even if the thresholds are known, the probability that a value will or will not be censored is not known. The censoring probability is given by:

$$P(\mathbf{x}_i'\beta \geqslant c) = P(\varepsilon < \mathbf{x}_i'\beta - c) = \Phi(\frac{\mathbf{x}_i'\beta - c}{\sigma}), \text{ since} \tag{2.9}$$

$$Y_i^* = \mathbf{x}_i'\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2) \tag{2.10}$$

Applying the logarithm to the likelihood function,

$$\ln(\mathscr{L}(y,d)) = \sum_{i=1}^{n}\{d_i \ln(f(y_i)) + (1-d_i)\ln(S_f(y_i))\} = \tag{2.11}$$

$$= \sum_{i=1}^{n} d_i \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2}\right) + \sum_{i=1}^{n}(1-d_i)\ln\left(\Phi(\frac{\mathbf{x}_i'\beta - c}{\sigma})\right)$$

By taking the derivative of the logarithm of the likelihood in order to the parameters β, $(j = 1, ..., p)$,

$$\frac{\partial \ln(L(y,d))}{\partial \beta_j} = \sum_{i=1}^{n}(1-d_i)\frac{\phi(\frac{\mathbf{x'}_i\beta - c}{\sigma})\frac{x_{ij}}{\sigma}}{\Phi(\frac{\mathbf{x'}_i\beta - c}{\sigma})},\ j = 1, ..., p \tag{2.12}$$

$$\frac{\partial \ln(L(y,d))}{\partial \beta_0} = \sum_{i=1}^{n}(1-d_i)\frac{\phi(\frac{\beta_0 - c}{\sigma})\frac{1}{\sigma}}{\Phi(\frac{\beta_0 - c}{\sigma})} \tag{2.13}$$

The next step would be to set these derivatives equal to zero to obtain the parameter estimates. Given the complexity of the expressions, analytical solutions are not available and therefore we resort to the EM Algorithm.

2.4 Logistic Regression Models

The difference between linear regression and logistic regression models is found in the response variable. In logistic regression the outcome is binary or dichotomous. Thus, the outcome only takes value 0 (absence/no) or 1 (presence/yes). This type of response variable requires the use of a model that estimates the probability of a specific event, taking into account the explanatory variables. The aim is to find the most suitable and most parsimonious model that describes the relationship between the response variable and the covariates. (Bagley et al., 2001)

As the response variable is binary,

$$Y_i \sim Bernoulli(p_i)\ ,\ \ i = 1, ..., n \tag{2.14}$$

Therefore,

$$E(Y_i) = P(Y_i = 1) = p_i,\ p_i \text{ is the probability of the success occurring.} \tag{2.15}$$

$$Var(Y_i) = p_i(1-p_i) \tag{2.16}$$

To model $E(Y_i) = p_i$ a link function between p_i and $\mathbf{x}_i$ is needed such that values on the right-hand side of the equation can be assumed on the left-hand side. Using the logit link function:

$$\text{logit}(p_i) = \ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,\ \ k = 1, ..., p+1 \tag{2.17}$$

The logit may range from $-\infty$ to $+\infty$ depending on the range of $\mathbf{x}$.

$$p = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_k x_k)}} \tag{2.18}$$

In linear regression models, $\varepsilon \sim N(0, \sigma^2)$. This is another difference between these two models. In logistic regression models, ε can take two values:

$$\varepsilon = \begin{cases} 1-p & \text{with prob } p \quad \text{if } y = 1 \\ -p & \text{with prob } 1-p \quad \text{if } y = 0 \end{cases}$$

Thus, ε has a distribution with mean zero and variance $p(1-p)$

It is necessary to estimate the unknown parameters to adjust the logistic regression model. The estimation is done via maximum likelihood method.

For any pair $(\mathbf{x_i}, y_i)$ the contribution to the likelihood function is p_i when $y_i = 1$, because $P(Y = 1|x) = p$, and is $1 - p_i$ when $y_i = 0$, because $P(Y = 0|x) = 1 - p$. Therefore, the likelihood function is as follows (Hosmer and Lemeshow, 2000):

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}. \tag{2.19}$$

Applying the logarithm to the likelihood function,

$$\ln(\mathscr{L}(\beta)) = \sum_{i=1}^{n} \{y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\} = \tag{2.20}$$

$$\sum_{i=1}^{n} \{y_i \ln(p_i) + \ln(1-p_i) - y_i \ln(1-p_i)\} =$$

$$\sum_{i=1}^{n} \{\ln(1-p_i) + y_i \ln \tfrac{p_i}{1-p_i}\}$$

Finally, the partial derivatives in order to the parameters are determined and made equal to 0 to find the estimates of the parameters:

$$\sum_{i=1}^{n} x_{i0}(y_i - p_i) = 0 \Leftrightarrow \sum_{i=1}^{n} (y_i - p_i) = 0 \tag{2.21}$$

$$\sum_{i=1}^{n} x_{im}(y_i - p_i) = 0 \ , \mathrm{m} = 1, ..., \mathrm{k} \tag{2.22}$$

The equation 2.21 is for determining the estimate of the β_0 parameter, while the equation 2.22 is for the remaining parameters of the model equation.

Interpreting the results

Odds

Odds is expressed as the quotient of the probability of the event occurring and the probability of the event not occurring.

$$\text{odds} = \frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}, \tag{2.23}$$

where is p is the probability of the event to ocurr.

Odds Ratio

Odds Ratio (OR) represents the odds of an outcome occurring in the presence of a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} \tag{2.24}$$

Regarding OR, when the parameter is positive, the exponential is greater than 1, which means that there is an increase in the chances of the occurrence of success. When the parameter is negative, the exponential is less than 1, so there is a decrease in the chances of the occurrence of success. Therefore, positive parameters favour the event and negative parameters protect the event.

Interpretations vary slightly depending on the type of variable.

Quantitative variable

The e^{β_j} corresponds to the OR associated with a unit increase in variable x_j.
In practical terms, if $\beta_j > 0$, then $(e^{\beta_j} - 1) \times 100\%$ represents the increase in the chance of success for each unit more in the predictor variable. If $\beta_j < 0$ then $(1 - e^{\beta_j}) \times 100\%$ represents the decrease in the chance of success for each unit more in the predictor variable.

Categorical variable

- Categorical binary
  As the variable is dichotomous, it has only one parameter associated and that parameter duly exponentiated is OR. The interpretation is similar to the interpretation for quantitative variables. If $\beta_j > 0$, then $(e^{\beta_j} - 1) \times 100$ represents the increase in the ratio of chances. If $\beta_j < 0$ then $(1 - e^{\beta_j}) \times 100$ represents the decrease in the ratio of chances.
- Categorical with more than 2 categories
  The interpretation of the parameters associated with the various levels of this variable is similar to the interpretation for the binary variable always compared with the reference category.

2.5 Poisson Regression Models

The poisson distribution is the most simple distribution for modeling counting data, such as the number of event occurrences during a particular time period.

As the response variable is a counting,

$$Y_i \sim Poisson(\mu_i) \ , \ \ i = 1, ..., n \tag{2.25}$$

Therefore,

$$E(Y_i) = \mu_i = Var(Y_i) \tag{2.26}$$

The linear regression model cannot be used since λ only takes positive values. One option to solve this problem would be a logarithmic transformation. Therefore, for these models the link function used is the logarithm. In a scenario where there are k independent variables, the equation of the model is as follows:

$$\log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i \tag{2.27}$$

The parameters are estimated through maximum likelihood. The maximum likelihood method produces values for the unknown parameters that maximize the probability of obtaining the observed data set.

The contribution of each observation to the likelihood function is $P(Y_i = y_i|\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$. So, the likelihood function is as follow (Autumn, 2016, Notes et al., 2015):

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \tag{2.28}$$

Applying the logarithm to the likelihood function,

$$\ln(\mathscr{L}(\beta)) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \mu_i - \ln(y_i!) = \tag{2.29}$$

$$= \sum_{i=1}^{n} \{y_i \sum_{j=0}^{k} \beta_j x_{ij} - \exp(\sum_{j=0}^{k} \beta_j x_{ij}) - \ln(y_i!)\}$$

Finally, it is necessary to determine the partial derivatives in order to the parameters and equal these expressions to zero to find the estimates of the parameters.

$$\frac{\partial \ln(\mathscr{L}(\beta))}{\partial(\beta_m)} = 0 \Leftrightarrow \sum_{i=1}^{n} x_{im}(y_i - \exp(\sum_{j=0}^{k} \beta_j x_{ij})) = 0, \;\; m = 0, ..., k \tag{2.30}$$

This type of model may have a detail, since subjects may not all have the same exposure time and since exposure time affects the response, the interest becomes in modelling the λ. In order to control the exposure time an offset is used to mark the time interval. The offset term is a "structural" predictor. Its coefficient is not estimated by the model being assumed to take the value 1. The offset values are simply added to the linear predictor of the response variable. This is especially useful in Poisson regression models, where each case may have different levels of exposure to the event of interest.

With the introduction of the offset term,

$$\lambda_i = \frac{E(Y|x_i)}{E_i} = \frac{\mu_i}{E_i}, \tag{2.31}$$

where E_i is the exposure,

$$\log(\lambda_i) = \log(\frac{\mu_i}{E_i}) = \log(\mu_i) - \log(E_i) \tag{2.32}$$

Thus, the model with the introduction of the offset is written as follows:

$$\log(\mu) = \log(E) + \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i \tag{2.33}$$

2.6 Generalized Linear Models - Longitudinal Data

Longitudinal data are data collected for the same individual at several different time points. When modelling longitudinal data by the normal linear model, the estimates of the regression coefficients remain consistent. However, these are inefficient because the variance of the estimated regression coefficients is biased. When there is more than one observation per individual, the issue of independence between observations is compromised. One way to solve the problem is to introduce random effects into the model. It is assumed that the response is a linear function of the explanatory variables with regression coefficients that vary across individuals. The random coefficients explain the dependence between responses from the same individual.

To use a longitudinal approach, the following could be used Generalized Linear Mixed-Effects Models (Appendice A) or Generalized Estimating Equation Models.

2.7 Generalized Estimating Equations

Generalized linear models are not applicable to longitudinal data due to the fact that the assumption of independence of observations fails immediately. To circumvent this problem, in addition to the mixed effects models approach one can also use the Generalized Estimating Equations models.

Generalized Estimating Equations (GEE) is a method of modelling longitudinal or clustered data. They are commonly used with non-normal data, such as binary data or count data. The name refers to a set of equations that are solved to obtain parameter estimates, i.e. the model coefficients (Ballinger, 2004).

Mixed effects models are generally referred to as conditional models, because they allow to estimate different parameters for each subject or cluster. That is, the parameter estimates are conditional on the subject or group. Therefore, it is possible to able to understand the variability between subjects or groups and can obtain a population model, but which is only an average of the subject specific models. On the other hand, GEE are called marginal models, due to the fact that they only estimate the global mean. It is introduced in the construction of the model the information that there is a dependency structure between the observations so that the standard error of the parameters is properly estimated and not underestimated.

The response vector for subject i $(i = 1,...,n)$ is

$$Y_i = (Y_{i1},...,Y_{ir}), \tag{2.34}$$

where r is the number of observations of subject i.

As the Generalized Estimating Equations are referred to as marginal models because they only estimate the overall mean, then:

$$\mu_{ij} = X_{ij}\beta,\ i = 1,...,n,\ j = 1,...,t \tag{2.35}$$

The format of the equation of GEE model is as follows:

$$Y_{ij} = \beta_0 + \sum_k X_{ijk}\beta_k + Corr + \varepsilon_{ijk}, \tag{2.36}$$

where y_{ij} is the outcome on subject i at moment j.

Depending on the context of the problem, the model is adapted according to the response variable.

In estimating these models, a naive linear regression analysis is first elaborated, but observations within subjects are assumed to be independent. Then, residuals are obtained from the first model and a correlation matrix is estimated from these residuals. Subsequently, the regression coefficients are readjusted thus correcting for correlation. The correlation structure between subjects is treated as a covariate.

Various types of correlation structures can be used in GEE models. When estimating the model, one type of correlation for the repeated measures has to be assumed. The correlation types are (Hardin and Hilbe, 2002):

- Independence $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

  Dependency structure that is used in linear regression models.

- Exchangeable $\begin{bmatrix} 1 & p & p \\ p & 1 & p \\ p & p & 1 \end{bmatrix}$

  This type of dependency structure is used when there is no time dependency and when any type of permutation of the repeated measures is valid.

- Autoregressive $\begin{bmatrix} 1 & p & p^2 \\ p & 1 & p \\ p^2 & p & 1 \end{bmatrix}$

  Dependency structure that assumes a time dependency between observations, if within each group there is a natural order of observations.

- Unstructured $\begin{bmatrix} 1 & p_1 & p_2 \\ p_1 & 1 & p_3 \\ p_2 & p_3 & 1 \end{bmatrix}$

  This dependency structure imposes no structure on the correlation matrix.

2.8 Model diagnosis for GLMs

2.8.1 Assumptions

To evaluate the selected model, a residual analysis has to be performed and there are assumptions that cannot be violated:

- $Y_1, Y_2, ..., Y_n$ are independent.
- The response variable (Y_i) need not follow a normal distribution, but normally assume a distribution of the exponential family.
- These models do not assume a linear relationship between the response variable and the explanatory variables, but rather assume a linear relationship between the expected response transformed in terms of the link function and the explanatory variables.
- Homogeneity of variance need not be satisfied. It is not even possible in many cases, given the structure of the model, as in the case of the poisson distribution where $E(Y) = Var(Y)$.
- The errors have to be independent but need not be normally distributed.

2.8.2 Goodness-of-fit

Deviance test

There are several ways to compare the models. When the models are nested, the analysis of deviance is used through the likelihood ratio test. Given two models M_s and M_f, involving respectively, s and f parameters $(s < f)$, M_s is said to be nested in M_f $(M_s \subset M_f)$, if all the parameters present in model M_s are present in model M_f

H_0: The variables that are present in the M_f model but not present in the M_s model are all irrelevant for modelling Y.

vs

H_1: At least one of those variables is relevant for modelling Y.

Test statistics:

$$2(\log(L_f) - \log(L_s)) \sim \chi^2_{p_s - p_f} \text{ under } H_0, \tag{2.37}$$

where
L_s is the maximized likelihood under the M_s model and
L_f is the maximized likelihood under the M_f model.

The closer the estimated model is to the data the lower the value of the deviation function will be.

Hosmer-Lemeshow test

In addition to the deviance test, there is also a test to check the quality of the adjustment to the data, specific to models with a binary response variable.

The Hosmer-Lemeshow goodness-of-fit test is based on dividing the sample up according to the predicted probabilities according to $1 - \hat{p}$ (2.18) (Fagerland and Hosmer, 2012).

Afterwards, according to the calculated probabilities, g groups are formed, such that each group contains approximately $\frac{n}{g}$ observations. The hypotheses under study:

H_0 : The fitted model is the correct model *vs* H_1 : The fitted model is not the correct model

Test statistic: overline

$$C = \sum_{k=1}^{g} \frac{(O_k - n'_k \overline{\pi}_k)^2}{n'_k \overline{\pi}_k (1 - \overline{\pi}_k)} \sim \chi^2_{g-2} \text{ under } H_0, \tag{2.38}$$

where, n'_k is the total number of subjects in the k^{th} group, O_k is the number of successes among the elements whose estimated probability is in group k, and $\overline{\pi}_k$ is the average of the estimated probabilities in group k.

Once the test is concluded, the hypothesis that the model is correct is rejected when $C > \chi^2_{(1-\alpha,g-2)}$.

2.8.3 Residuals

The raw or response residuals are the deviations between the observed values y_i and the adjusted values $\hat{y}_i$ (Pearson and Pearson, n.d.).

$$e_i = y_i - \hat{y}_i, \;\; i = 1, ..., n \tag{2.39}$$

Some types of residuals are presented below and their calculation differs depending on the type of model being used. The formulas presented are the same for Poisson and Binomial models, with the proper adjustments. For Binomial models $\hat{\mu}_i = \hat{p}_i$.

2.8.3.1 Pearson's residual

The Pearson residuals are the raw residuals divided by the estimated standard error of observed values.

$$r_{pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \tag{2.40}$$

The variance function depends on the type of model being applied. In the case of Binomial models $V = \hat{\mu}_i(1 - \hat{\mu}_i)$, in the case of Poisson models $V = \hat{\mu}_i$.

2.8.3.2 Standardized Pearson's residual

As the obtained pearson residuals may not have a unit variance, the residuals are standardized. The raw residuals are standardized by their estimated standard errors

$$r'_{pi} = \frac{pr_i}{\sqrt{\hat{\phi}(1 - h_i)}}, \tag{2.41}$$

where, h_i are the diagonal values of the hat matrix and represents the distance between the i^{th} observation in relation to the remaining observations versus its order number and ϕ is 1 for Binomial and Poisson models.

2.8.3.3 Deviance Residuals

Alternatively to the Pearson standardized residuals, one can also use the deviance residuals. The deviance residuals measures the disagreement between one of the fitted log-likelihood components and the corresponding log-likelihood component obtained if each point were adjusted exactly.

$$r_{di} = signal(y_i - \hat{\mu}_i)\left\{2y_i \ln(\frac{y_i}{\hat{\mu}_i}) + 2(n_i - y_i)\ln(\frac{n_i - y_i}{n_i - \hat{\mu}_i})\right\}^{\frac{1}{2}} \tag{2.42}$$

2.8.4 Discordant observations

Discordant observations are divided into outliers and/or influential observations.

Outliers

The test for outliers produces studentized residuals $E_1^*, E_2^*, ..., E_n^*$, with interest falling on the largest absolute $E_i^* = E_{max}^*$. To avoid the problems of simultaneous inference, the p-value with bonferroni adjustment is used for E_i^*. The p-value with bonferroni adjustment for testing the statistical significance of E_{max}^* is $p = 2np'$, where $p' = P(t_{n-k-2} > E_{max}^*)$, so this test will detect extreme outliers (Sugawara and Nikaido, 2014). Which tests the following hypotheses:

$$H_0 : \text{There are no outliers vs } H_1 : \text{There are outliers}$$

Influential observations

For existence of influential values Cook's distance plot is used. Cook's distance can be calculated from the following expression (Fox and Weisberg, 2011):

$$D_i = \frac{e_{Si}^2}{k+1} \times \frac{h_{ii}}{1-h_{ii}}, \tag{2.43}$$

where, e_{Si}^2 is the squared standardized residual, k is the number of coefficients in the regression model and h_{ii} are the diagonal values of the hat matrix.

The following rules are used to determine whether an observation is influential (University, 2018, Cabral, 2019):

- If D_i is greater than 0.5, the observations corresponding to that distance can be an influential observation, so this observation is worth investigating further.
- If D_i is greater than 1, the observation corresponding to that distance is very likely to be an influential observation.

2.9 Model diagnosis for Tobit model

The analysis of residuals for Tobit models is very similar to the analysis mentioned in section 2.8. The response variable needs to be censored quantitative continuous.

Influential observations

To check for the existence of influential observations, the dfbetas for the intercept and for each parameter associated to the explanatory variables (STATA, 2013).

$$Dfbeta_k = \frac{r_k u_k}{\sqrt{U^2(1-h_k)}},\ k = 1, ..., p\ , \tag{2.44}$$

where,
r_k are the studentized residuals,
u_k are the residuals obtained by regressing x_i on the remaining regressors,
$U^2 = \sum_{k}^{p} u_k^2$,
h_k are the diagonal value of the hat matrix corresponding to each regression variable.

The threshold will be $\pm\frac{2}{\sqrt{n}}$. All observations that exceed this ceiling will be considered influential and should be dropped from the model.

2.10 Statistical packages

For the calculation of prevalence, the package **survey** (Lumley, 2021) was used to use the **svydesign** and **svyciprop** functions. For the elaboration of Tobit models with a cross-sectional perspective, the **AER** package (Functions and Kleiber, 2022) was used so that the **tobit** function could be used and the **evd** package (Stephenson, 2022) was used to produce the graphs referring to these models. Whereas, for the longitudinal perspective of the Tobit models it was not possible to perform this analysis in the R software and therefore the Stata software was used and the **metobit** function was used. For the GEE models, the package **gee** (Generalized and Equation, 2022) was used, for the use of the **gee** function. For the elaboration of the graphs of this analysis the package **car** (John et al., 2022) was used. Throughout the project it was also used the package **ggplot2** (Create et al., 2022) to produce some graphics.

3. The Data

With the aim of assessing the health status of the Portuguese adult population, questionnaires were applied following a pre defined sampling design. The primary aim of the EpiDoC cohort is to examine the health determinants and outcomes of chronic non-communicable diseases and their impact on health care resource consumption. The individuals who answered the questionnaires were randomly selected. Candidates for participation were visited in their homes by a team of trained interviewers. The locations were selected using the primary sampling unit (PSU) according to the 2001 Census. The selected households and their addresses were identified using a random selection of points on the map of each location, where the interviewer initiated a systematic step count (defined for each location based on its size). Each selected household was visited, without prior contact, up to three times (including evenings and weekends) if the candidate was not present at previous visits. The eligible participants had to be over 18 years of age, not institutionalised, and among several household members, the individual selected had their birthday the closest to the interview date.

3.1 Data description

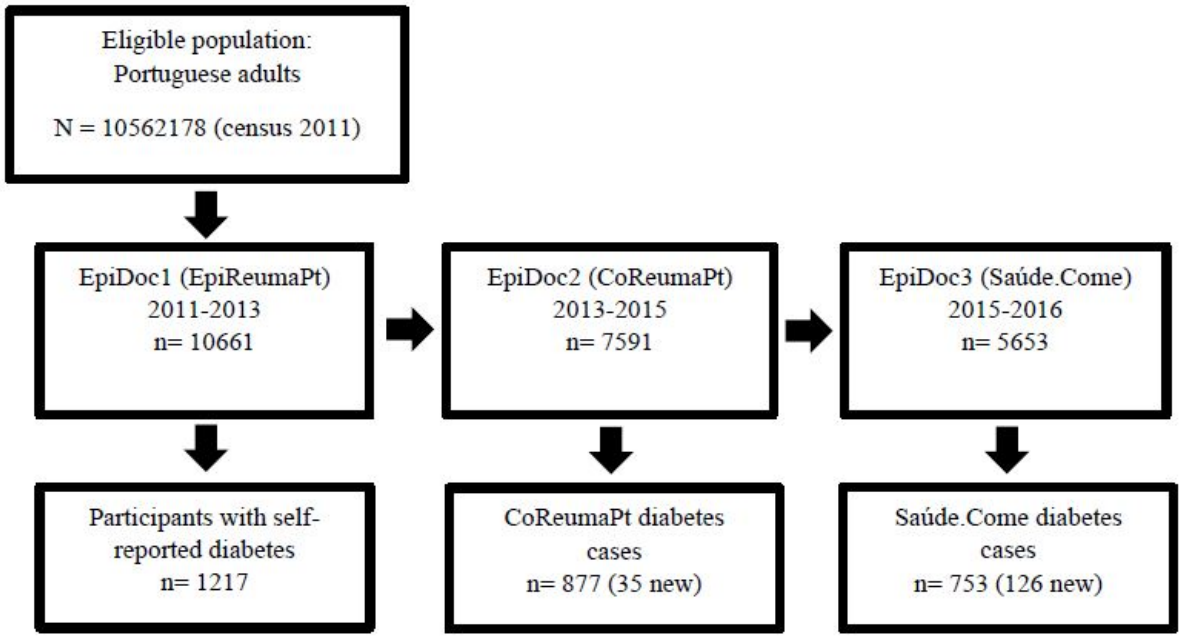

Fig. 3.1: Data flowchart.

For this analysis, data from the EpiDoc cohort (Dias et al., 2018) were used. There were three moments of data collection. The first moment of data collection of this cohort study was called EpiDoc1

(EpiReumaPt) and its main objective was the study of rheumatic diseases, data collection occurred between September 2011 and December 2013 and included 10661 individuals. Face-to-face interviews were conducted with the selected individuals. Of these 10661, only 3877 participated in a detailed clinical evaluation of RMD (Rheumatic and Musculoskeletal Diseases) performed by rheumatologists. The second wave, named EpiDoc2(CoReumaPt), took place between March 2013 and July 2015 and it involved 10153 individuals considered as eligible out of the 10661 individuals regarding EpiDoc1. However, only 7591 of these representing the adult Portuguese population, participated. In this ocasion, telephone interviews were conducted. This wave allowed for longitudinal data analysis, as well as adding new questions on lifestyle, health innovation and social interactions, socio-demographic and socio-economic questions, anthropometric measures, and so-forth. The third wave, EpiDoc3 (Health.Come), was conducted between September 2015 and July 2016. 9023 individuals were considered eligible, but only 5653 individuals participated in the study. Telephone interviews were again conducted and, as in the previous wave, data continued to be collected for a longitudinal analysis and new data was collected characterising household food insecurity and its consequences for the health of the population residing in Portugal.

As mentioned earlier, this project focuses mainly on diabetes. To characterise diabetes, only the first wave was used. Since diabetes is a chronic disease, once someone is diagnosed with the disease, it remains in the future.

For this purpose variables from the database were used and others were created through already existing variables.

- Sex: female, male.
- Age group: 18-25, 26-35, 26-45, 46-55, 56-65, 66-75, 76-85, $\geqslant$ 86.
- NUTS II: Norte, Centro, Lisboa, Alentejo, Algarve, Açores, Madeira.
- Ethnicity: caucasian, other.
- Marital status: single, married, divorced, widow(er), consensual union.
- Education level: 0, 1-4, 5-9, 10-12, $>$12.
- Employment status: - employed (Employed full-time, Employed part-time, Domestic worker,Temporally work disabled); - not employed (Unemployed, Student); - retired
- Lifestyle habits: - alcohol intake(Yes(Daily,Occasionally), No);
  - smoking habit (Yes(Daily,Occasionally), No); - exercise practice (Yes,No)
- Presence/absence of disease: diabetes, cholesterol, cardiac, high blood pressure, mental, allergies, rheumatic, gastrointestinal, pulmonary)
- BMI: underweight, normal, overweight, obese.
- Score EQ-5D.
- Occurrence of hospitalizations.
- Number of medical appointments.

- Wave.
- Time of exposure.

The variables were analysed in a cross-sectional and a longitudinal format.
Note that the education level, age and BMI variables are used in both a continuous and a categorical perspective.

The following variables were created: wave - in order to have an identifier for each wave of data collection in longitudinal format; Time of exposure - is expressed in months, assuming the value 12 for the first data collection, in the second data collection it assumes the value of the difference between the date of the interview of the second wave with the first wave and in the third data collection it assumes the value of the difference between the date of the interview of the third wave with that of the first wave; Number of medical appointments - corresponds to the sum of medical appointments in the private and public sector. Note that for chronic diseases, all persons who responded having the disease at wave n-1, at wave n would necessarily have the disease, so the inconsistent responses, where an individual had the disease and stopped having it at the next wave, were considered wrong and it was considered that the persons also had the disease at the next wave.

3.2 Descriptive Analysis

Table 3.1 contains the distribution and weighted prevalences of diabetics and non-diabetics in some characteristics. Weighted prevalences were calculated according to a constructed weighting that took into account the NUTS II, sex and age of each selected individual. Therefore, each individual represents a different weight for the analysis. In addition, the chi-square test was used to analyse the association between being or not being diabetic taking into account the remaining variables. The p-value associated with the chi-square test is also mentioned. The first wave reported that 1217 (8.3%) of the individuals were diabetic, and the remaining 9370 (91.7%) were non-diabetic. For the elaboration of the methodologies made in this chapter, only the data from the first collection was considered.

Table 3.1: Absolute frequencies and weighted prevalences referring to diabetics and non-diabetics.

		Diabetics n=1217	Non--Diabetics n=9370	p-value [1]
Sex	Male	448 (42.9)	3635 (47.7)	0.1918
	Female	769 (57.1)	5735 (52.3)	
Age group	18-25	10 (2.1)	816 (17.0)	«0.001
	26-35	16 (2.1)	1230 (21.0)	
	36-45	48 (5.6)	1778 (18.8)	
	46-55	138 (14.0)	1733 (16.3)	
	56-65	302 (26.3)	1733 (16.3)	
	66-75	391 (29.4)	1306 (9.3)	
	76-85	271 (18.0)	800 (5.1)	
	$\geqslant$ 86	41 (2.4)	131 (0.7)	
NUTS II	Norte	370 (34.2)	2729 (35.1)	«0.001
	Centro	249 (25.9)	1736 (22.6)	
	Lisboa	218 (22.8)	2252 (26.8)	
	Alentejo	92 (9.4)	574 (7.1)	
	Algarve	40 (3.1)	309 (3.8)	
	Açores	126 (2.4)	892 (2.2)	
	Madeira	122 (2.2)	878 (2.3)	
Ethnicity	Caucasian	1196 (96.1)	9073 (95.9)	0.0039
	Other	17 (3.9)	269 (4.1)	
Marital status	Single	73 (7.1)	1853 (31.4)	«0.001
	Married	746 (63.2)	5325 (49.2)	
	Divorced	52 (4.7)	754 (7.7)	
	Widow(er)	333 (23.3)	1063 (6.7)	
	Consensual union	13 (1.7)	366 (5.1)	
Education level	0 years	316 (21.5)	839 (5.4)	«0.001
	1-4 years	605 (50.1)	2912 (23.9)	
	5-9 years	156 (14.3)	2009 (23.5)	
	10-12 years	82 (9.7)	1831 (25.2)	
	>12 years	45 (4.5)	1717 (22.0)	
Employment status	Employed	297 (26.5)	4870 (55.6)	«0.001
	Not employed	77 (8.1)	1545 (23.1)	
	Retired	831 (65.4)	2885 (21.3)	
Lifestyle habits (Yes)	Alcohol intake	549 (51.3)	3925 (36.3)	«0.001
	Smoking habit	99 (10.7)	1989 (27.0)	«0.001
	Exercise practice	256 (22.3)	3226 (38.5)	«0.001
Diseases (Yes)	Cholesterol	706 (55.7)	2626 (21.6)	«0.001
	Cardiac	358 (30.4)	994 (8.4)	«0.001
	High blood pressure	797 (65.6)	2540 (19.1)	«0.001
	Mental	267 (21.8)	1340 (12.1)	«0.001
	Allergies	280 (21.8)	1994 (21.3)	0.1638
	Rheumatic	587 (46.5)	2378 (18.7)	«0.001
	Gastrointestinal	314 (26.1)	1506 (13.9)	«0.001
	Pulmonary	118 (9.6)	513 (5.0)	«0.001
BMI	Underweight	7 (0.9)	158 (2.4)	«0.001
	Normal	202 (19.8)	3839 (48.1)	
	Overweight	454 (37.8)	3324 (34.7)	
	Obese	445 (41.5)	1611 (14.9)	

[1] Qui-square test.

- Sex

There is a higher proportion of women in both categories, with 57.1% of diabetics and 52.3% of non-diabetics being women. However, there are no significant differences between diabetics and non-diabetics taking gender into account.

From another perspective, the prevalence of diabetes in women is 8.98% and in men 7.52%, so the prevalence of diabetes is similar in both sexes.

According to a report by Acta Médica Portuguesa, there was also a similar prevalence of diabetes for

both sexes, thus there is agreement with these results. (Santos et al., 2017).
Age does not seem to be a confounding factor as the average age is 67 years and 68 years for diabetic men and women, respectively.

- Age

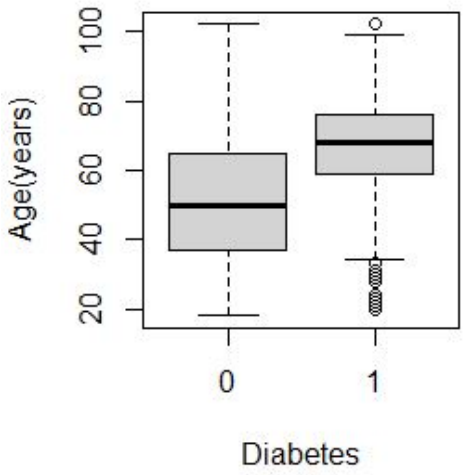

Fig. 3.2: Boxplot age in years for non diabetics(0) and diabetics(1).

According to Figure 3.2, the median age of diabetics is higher than the median age of non-diabetics. In diabetics, the distribution presents a slightly skewed pattern to the left.

The average age of diabetics at the date of the interview is 66.38 years. The most represented age group among diabetics is from 66 to 75 years (29.4%). As for non-diabetics, the most represented ages are lower than those of diabetics. Most non-diabetics are aged between 26 and 35 years (21%).
Looking at the data from another perspective, the prevalence of diabetes for each age group was 1.13%, 0.90%, 2.63%, 7.23%, 16.9%, 22.2%, 24.4% and 23.0%, respectively, from the first age group (18-25) to the last $\geqslant$86.Although the second and last age group departs from the pattern, the prevalence of diabetes increases with advancing age.
Also in the same report of the Acta Médica Portuguesa, it was concluded that the prevalence of diabetes increased significantly with age, which is in line with these results (Santos et al., 2017). There are significant differences between diabetics and non-diabetics taking age into consideration.

Below are the characteristics of each age group for diabetics:

Table 3.2: Sample characteristics of the age group variable for diabetics.

	Minimum	Mean	Standard Deviation	1st quartile	Median	3rd quartile	Maximum
18-25	20	22.2	1.40	21	22.5	23	24
26-35	28	31.56	2.39	29.75	32	33.25	35
36-45	36	41.35	2.90	39.75	41.50	44	45
46-55	46	51.09	2.81	49	51	54	55
56-65	56	60.84	2.75	59	61	63	65
66-75	66	70.38	2.73	68	70	73	75
76-85	76	79.28	2.56	77	79	81	85
>85	86	88.76	3.23	87	88	89	102

- NUTS II

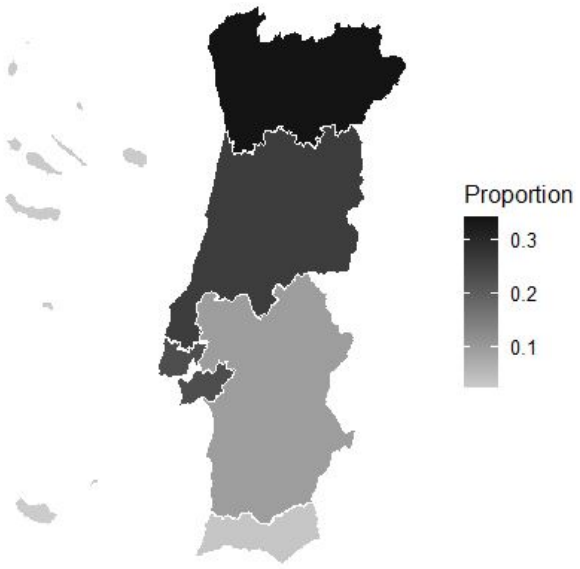

Fig. 3.3: Map of portugal divided by NUTS II and classified by weighted prevalence of diabetics.

Both in Table 3.1 and in the Figure 3.3 it can be observed that diabetics are more represented in the North region (34.2%). However, there are significant differences between diabetics and non-diabetics taking regions (NUTS II) into account.

- Ethnicity

The most abundant ethnicity is caucasian, in both diabetics (96.1%) and non-diabetics (95.9%) groups. Nevertheless, there are significant differences between diabetics and non-diabetics in relation to ethnicity. This was an expected result, since the sample was collected in Portugal, a country where the caucasian ethnicity is predominant.

- Marital Status

There were more married participants, in both categories, were 63.2% diabetics and 49.2% non-diabetics married. There are significant differences between diabetics and non-diabetics taking marital status into account.

- Education level

As for the levels of education, the highest proportion of diabetics is found in individuals who studied between 1 and 4 years (50.1%). Non-diabetics are mostly more educated. The largest proportion is from 10 to 12 years (25.2%).

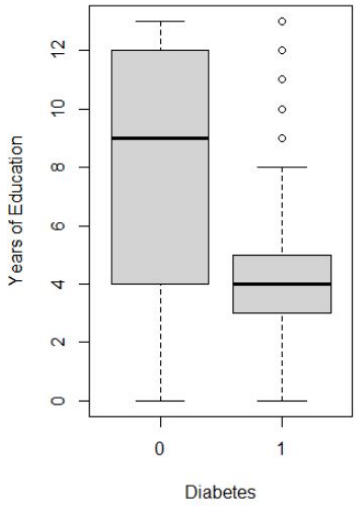

Fig. 3.4: Boxplots by the years of education in non diabetics and diabetics, respectively.

Analysing the boxplots, it is visible that the years of education of diabetics are lower than that of non-diabetics. Ignoring the outliers, it can be seen that the highest value of education for diabetics is below the median education of non-diabetics, that is, diabetics have less schooling. Age and education are related, as older people have a lower level of education. Since diabetes is associated with higher age, it is possible that age is considered a confounding factor for education.

Table 3.3: Sample characteristics of the variable age by education levels for diabetics.

	Minimum	Mean	Standard Deviation	1st quartile	Median	3rd quartile	Maximum
0	36	73.65	9.08	69	75	79	99
1-4	24	66.63	10.03	60	67	74	102
5-9	23	59.65	12.91	50.75	60	70	89
10-12	20	52.26	15.76	40.25	54	64	86
>12	22	57.11	16.70	46	59	68	85

By analysing Table 3.3, it can be observed that age increased with the decrease levels of education. That is, as mentioned before, it is possible to confirm that these two variables are inversely associated. Therefore taking into account this study, older people had less scholling. However, it is not to say that less schooling leads to diabetes.

- Employment Status

In regards to employment status, the following categories were considered: Employed, that includes employed full-time, employed part-time, domestic worker, temporally work disabled and working student; Not employed which includes unemployed, student, doensn't work, but lives on revenus and looking for the first job; Retired.

Through the results, the most common employment status in diabetics is retired (65.4%). However, age is a confounding factor for employment status, since, as mentioned above, most diabetics are older, which is also true for retired people. Still, according to an article in the American Diabetes Association, *"diabetes affects patients, employers, and society not only by reducing employment but also by contributing to work loss and health-related work limitations for those who remain employed."* (Tunceli et al., 2005).

- Lifestyle habits

A healthy lifestyle reduces the risk of developing diabetes. The lifestyle habits considered were alcohol intake, habit of smoking and practice of exercise.

Regarding alcohol intake, most diabetics stated that they drink alcohol, daily or occasionally (51.3%). On the contrary, most non-diabetics do not drink alcohol.As can be seen from Table 3.1 and as seen in Figure 3.5. Excessive alcohol consumption increases the risk of developing chronic inflammation of the pancreas. This condition leads to irreversible damage including the destruction of the cells responsible for producing, storing and releasing insulin. One of the consequences of this inflammation is a reduced ability of the pancreas to secrete insulin, which can potentially lead to type-2 diabetes (Mosel, 2022).

Fig. 3.5: Mosaic plot representing the distribution of diabetes by alcohol intake.

Regarding smoking and physical exercise, both the majority of diabetics and the majority of non-diabetics claim that they do not smoke or exercise.

The centre of disease control and prevention has found that smokers are 30-40% more likely to develop type 2 diabetes than non-smokers. And people with diabetes who smoke are more likely than non-smokers to have problems with their insulin dosage and the control of their disease (Centers for Disease Control and Prevention [CDCP], 2014). Knowing this and knowing that these patients have regular medical monitoring, patients are advised to stop smoking.

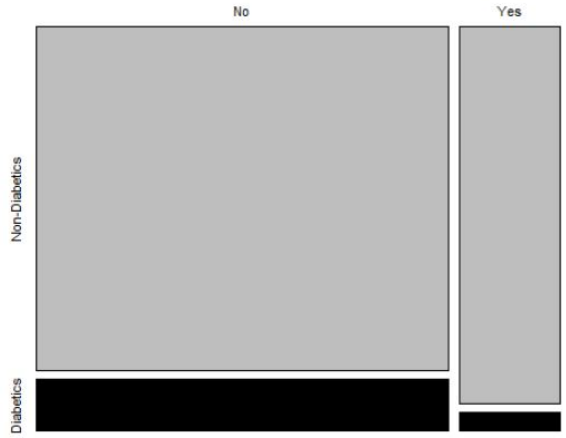

Fig. 3.6: Mosaic plot representing the distribution of diabetes by smoke habit.

Regarding the practice of physical exercise, the American Diabetes Association published an article saying that *"Exercise improves blood glucose control in type 2 diabetes, reduces cardiovascular risk factors, contributes to weight loss, and improves well-being. Regular exercise may prevent or delay type 2 diabetes development."* (Colberg et al., 2016). It would be expect that diabetics to exercise more, as it helps control blood glucose. 77.70% of diabetics do not exercise. Even so, non-diabetics should also exercise more, as it is a way of preventing this disease and many others.

Fig. 3.7: Mosaic plot representing the distribution of diabetes by excercise practice.

There are significant differences between diabetics and non-diabetics taking these lifestyle habits into account.

- Body Mass Index

Finally, in regards to body mass index (BMI), among diabetics the greatest proportion is obese (41.5%) and the majority of non-diabetics are of a weight considered normal (48.1%).

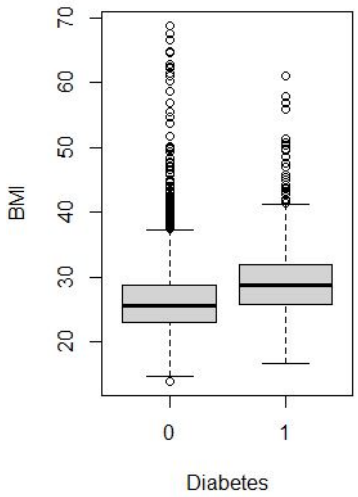

Fig. 3.8: Boxplot BMI in non-diabetics and diabetics, respectively.

The boxplot shows that the mean BMI is higher in diabetics. It is also clear that the mean BMI is approximately equal to the 3^{rd} quartile for non-diabetics. Even so, there are several outlier candidates and the highest value of BMI corresponds to non-diabetics, as well as the lowest value.
There are significant differences between diabetics and non-diabetics taking BMI into account.

- Diseases

The most represented comorbidity for diabetics, in terms of proportion, is high blood pressure (65.6%). As for non-diabetics, this proportion is higher for cholesterol (21.6%) and allergies (21.3%).
There are significant differences between diabetics and non-diabetics taking into account the presence or absence of each of the diseases, except for allergies where the difference is not significant.

3.2.1 Disease associations

In order to group diseases, association rules technique was used. Only diabetic individuals were considered so that it was possible to find out which other diseases were more present in individuals with diabetes. A total of 1099 diabetics were considered. The diseases considered in addition to diabetes were: Cholesterol, High blood pressure, Mental, Cardiac, Pulmonary, Allergies, Digestive and Rheumatic.

Table 3.4: Frequencies of disease groups in diabetic individuals

	Support	Confidence	Coverage	Count
{High blood pressure}	0.66	0.66	1.00	723
{Cholesterol}	0.59	0.59	1.00	647
{Rheumatic}	0.49	0.49	1.00	539
{High blood pressure} ⇒ {Rheumatic}	0.35	0.53	0.66	384
{Rheumatic} ⇒ {High blood pressure}	0.35	0.71	0.49	384
{Rheumatic} ⇒ {Cholesterol}	0.33	0.67	0.49	359
{Cholesterol} ⇒ {Rheumatic}	0.33	0.55	0.59	359
{High blood pressure} ⇒{Cholesterol}	0.44	0.67	0.66	488
{Cholesterol} ⇒{High blood pressure}	0.44	0.75	0.59	488
{Cholesterol, High blood pressure} ⇒ {Rheumatic}	0.26	0.58	0.44	282
{Cholesterol, Rheumatic} ⇒ {High blood pressure}	0.26	0.79	0.33	282
{High blood pressure, Rheumatic} ⇒ {Cholesterol}	0.26	0.73	0.35	282

The table contains the most frequent groups of diseases. The respective values correspond to the support representing the frequency of a disease or set of diseases. The confidence that represents the fact of having a certain disease knowing that one has a disease or a set of diseases.The coverage represents the frequency of the predecessor. And the count, number of diabetic individuals with that set of diseases. For this purpose, were only considered sets with a support greater or equal to 0.25 and a confidence greater or equal to 0.3.

Taking into account only the 1099 diabetics considered in this analysis, it is possible to verify that most of them have high blood pressure. Of these diabetics, 66% present high blood pressure, 59% had cholesterol and 49% had rheumatic disease. Considering the existence of two diseases, most of the diabetics are characterised as having high blood pressure and rheumatic diseases. It can be said that 53% of diabetics who had high blood pressure also had rheumatic disease. And on the other hand, 71% of diabetics who had rheumatic disease also had high blood pressure. In a set of three diseases, most diabetics had cholesterol, high blood pressure and rheumatic disease. It can be said that 58% of diabetics who had cholesterol and high blood pressure also had rheumatic disease.

3.2.2 Disease groups

Still with the objective of grouping the diseases using another type of analysis, it was resorted to the distance matrix and the elaboration of a dendrogram. The dendrogram is a tree diagram that shows the groups formed by grouping observations at each step and their similarity levels.

Table 3.5: Distance matrix

	Diabetes	High blood pressure	Cholesterol	Cardiac	Mental	Allergies	Gastrointestinal	Rheumatic	Pulmonary
Diabetes	0	14.6	15.7	2.3	6.6	13.5	7.1	13.4	3.3
High blood pressure	14.6	0	2.3	12.7	10.8	8.0	8.4	4.1	17.6
Cholesterol	15.7	2.3	0	13.7	11.4	7.5	9.2	4.9	18.6
Cardiac	2.3	12.7	13.7	0	5.4	11.6	5.3	11.7	5.1
Mental	6.6	10.8	11.4	5.4	0	8.5	3.0	8.6	8.6
Allergies	13.5	8.0	7.5	11.6	8.5	0	7.2	7.2	15.6
Gastrointestinal	7.1	8.4	9.2	5.3	3.0	7.2	0	6.9	9.8
Rheumatic	13.4	4.1	4.9	11.7	8.6	7.2	6.9	0	16.2
Pulmonary	3.3	17.6	18.6	5.1	8.6	15.6	9.8	16.2	0

The distance matrix was elaborated by calculating the prevalence of an individual having a particular disease and a particular characteristic. For this analysis, the characteristics taking into account were gender, age group and NUTS II. Which originated the distance matrix

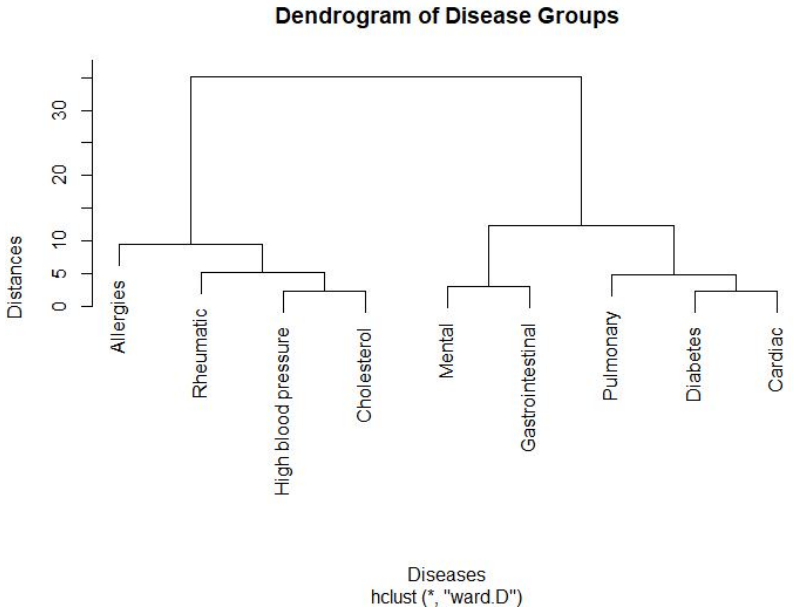

Fig. 3.9: Dendrogram of disease groups

This dendrogram was elaborated using the Ward method. Through the dendrogram and the distance matrix, the first diseases to join are diabetes and cardiac disease and with the same distance also joined high blood pressure and cholesterol. The pair mental and gastrointestinal diseases are then added to the first group mentioned. The group of high blood pressure and cholesterol is later joined by rheumatic disease and finally allergies join this group. So, the two major groups are made up of the following diseases:

- Cardiac, diabetes, pulmonary, gastrointestinal and mental.
- Cholesterol, high blood pressure, rheumatic and allergies.

4. Impact of diabetes on quality of life

The aim of the study is to assess diabetes, taking into account that it is affected by other factors and therefore other variables are included in the models. For the modelling of these data we have the following explanatory variables: Sex (reference class: Male); Age; NUTS II (reference class: Lisbon); BMI (reference class: Not overweight); Employment status (reference class: employees); Years of education; Diabetes; High blood presure; Cardiac; Mental; Gastrointestinal; Rheumatic; Allergies; Pulmonary; Alcohol intake; Smoking habit; Exercise practice. For all binary variables the reference class is "no", except for the variable sex.

4.1 EQ-5D-3L - Tobit Model

The EQ-5D-3L questionnaire was used to assess the impact of diabetes on quality of life. This questionnaire is an instrument for measuring health-related quality of life that allows generating an index representing the value of an individual's health status (EuroQol Research Foundation, 2018).

It started to be developed by the EuroQol group in 1987 and was made public in 1990. In this questionnaire there are 5 dimensions concerning mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each category has 3 possible answers, and a score is given for the type of answer. The score goes from 1 to 3, where level 1 corresponds to having no problems, level 2 corresponds to some problems and level 3 represents extreme problems. Therefore, this mechanism makes it possible to describe a total of $3^5 = 243$ distinct health states. At the end there will be a 5 digit code, for example, the code 11111 indicates that there are no problems in any of the 5 dimensions (Ferreira et al., 2013).

After ascertaining the 5 digit code, it is possible to transform it into an aggregated score of the 5 dimensions, referring to the state of health of each individual. This score admits values on a scale from 1 (perfect health) to -1, where the value 0 is considered as death and the negative values corresponding to states of health considered worse than death.

Table 4.1: Results of the EQ-5D-3L questionnaire application - 1^{st} wave

	Mobility (%)	Personal Care (%)	Usual Activities (%)	Pain/ Discomfort (%)	Anxiety/ Depression (%)	Mean Score
No problems	8228 (77.2)	9707 (91.1)	8653 (81.2)	6909 (64.9)	8351 (78.7)	
Some problems	2393 (22.5)	852 (8.0)	1845 (17.3)	3390 (31.8)	2062 (19.4)	0.79
Problems	35 (0.3)	96 (0.9)	154 (1.4)	354 (3.3)	204 (1.9)	

Table 4.2: Results of the EQ-5D-3L questionnaire application - 2^{nd} wave

	Mobility (%)	Personal Care (%)	Usual Activities (%)	Pain/ Discomfort (%)	Anxiety/ Depression (%)	Mean Score
No problems	4951 (67.4)	6046 (82.3)	5379 (73.2)	4116 (56.1)	4623 (63.5)	
Some problems	2346 (31.9)	1199 (16.3)	1882 (25.6)	2733 (37.3)	1971 (27.1)	0.70
Problems	47 (0.6)	99 (1.3)	86 (1.2)	485 (6.6)	664 (9.4)	

Table 4.3: Results of the EQ-5D-3L questionnaire application - 3^{rd} wave

	Mobility (%)	Personal Care (%)	Usual Activities (%)	Pain/ Discomfort (%)	Anxiety/ Depression (%)	Mean Score
No problems	4123 (74.6)	4578 (82.8)	4326 (78.2)	3426 (62.0)	3967 (72.5)	
Some problems	1378 (24.9)	889 (16.1)	1145 (20.7)	1809 (32.7)	1242 (22.7)	0.75
Problems	27 (0.5)	64 (1.2)	59 (1.1)	294 (5.3)	262 (4.8)	

In any of the waves, most individuals reveal absence of problems. The physical component is in a better condition than the pain and mental component, as the values of the categories mobility, personal care and usual activities are systematically higher than the others. Throughout the waves the scenario does not change much, which can also be seen by the average score. Even so, the highest mean score is in the first wave and so that is when the quality of life is best, on average.

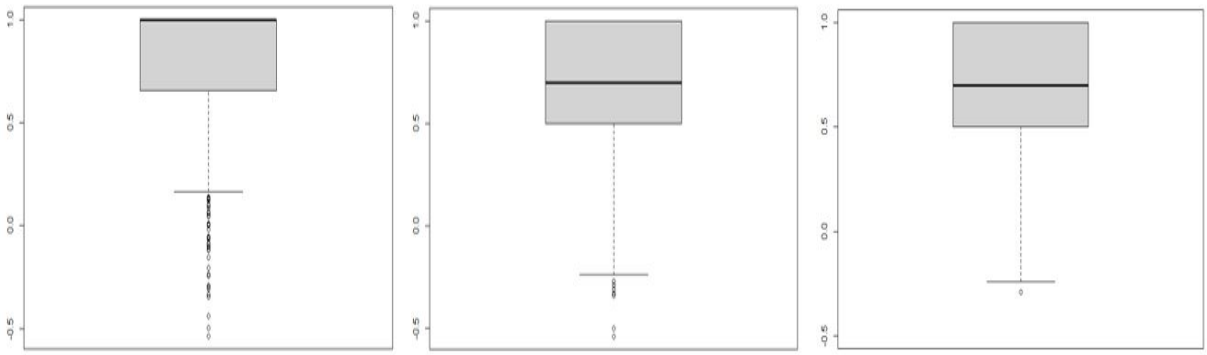

Fig. 4.1: Boxplots of the EQ-5D score for each wave.

The distribution of the quality of life score is practically equal in the 2^{nd} and 3^{rd} waves, but there is a bias to the right, which means that there is a higher concentration of lower score values. In the first wave, the median corresponds to the 3^{rd} quartile, with a bias to the left. Therefore, 50% of the values of the scores in 1^{st} wave are 1. Not taking into account the outliers, the score in the first wave does not reach 0, and in the remaining waves it registers negative values. From the boxplots analysis, it seems that the quality of life got worse after 1^{st} wave.

4.1.1 Modelling score quality of life in a transversal approach

In order to assess the impact of diabetes on quality of life, measured by the EuroQol Five Dimensional Questionnaire score, the quality of life score was modelled in each data collection wave, as well as in a longitudinal perspective, in order to assess the evolution of this measure over time. The variable quality of life score was modelled as a function of several explanatory variables, namely diabetes, using Tobit models.

Table 4.4: Results of the Tobit Regression models concerning the quality of life, for the three waves, respectively.

	1^{st} wave		2^{nd} wave		3^{rd} wave	
Variable	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value
Intercept	1.343	<<0.001 ***	1.449	<<0.001 ***	1.348	<<0.001 ***
Sex Female	-0.092	<<0.001***	-0.153	<<0.001***	-0.159	<<0.001***
Age	-0.003	<<0.001***	-0.004	<<0.001***	-0.002	0.009**
NUT II Norte	0.019	0.134	-0.030	0.036*	-0.037	0.152
NUT II Centro	0.002	0.906	-0.042	0.007**	-0.067	0.016*
NUT II Alentejo	0.046	0.029*	-0.013	0.604	-0.033	0.413
NUT II Algarve	0.032	0.247	-0.027	0.369	0.029	0.591
NUT II Açores	-0.013	0.447	-0.067	<<0.001***	-0.074	0.018*
NUT II Madeira	0.001	0.965	-0.058	0.002**	-0.068	0.035*
Employment Status Not employed	-0.057	<<0.001***	-0.047	0.006**	-0.091	0.003**
Employment Status Retired	-0.021	0.134	0.009	0.556	-0.008	0.734
Years of education	0.009	<<0.001***	0.015	<<0.001***	0.018	<<0.001***
Diabetes (Yes)	-0.043	0.004**	-0.013	0.416	-0.006	0.811
Cholesterol (Yes)	-0.033	0.002**	-0.032	0.004**	-0.008	0.649
Cardiac (Yes)	-0.076	<<0.001***	-0.058	<<0.001***	-0.067	0.004**
High blood pressure (Yes)	-0.025	0.025*	-0.029	0.015*	0.007	0.694
Mental (Yes)	-0.123	<<0.001***	-0.136	<<0.001***	-0.116	<<0.001***
Rheumatic (Yes)	-0.232	<<0.001***	-0.195	<<0.001***	-0.206	<<0.001***
Pulmonar (Yes)	-0.060	0.001**	-0.061	0.002**	-0.027	0.393
Gastrointestinal (Yes)	-0.089	<<0.001***	-0.078	<<0.001***	-0.068	0.002**
Allergies (Yes)	-0.038	<<0.001***	-0.019	0.112	-0.003	0.891
BMI	-0.004	<<0.001 ***	-0.010	<<0.001 ***	-0.008	<<0.001 ***
Alcool intake (Yes)	0.039	<<0.001***	0.036	<<0.001***	0.028	0.139
Smoking habit (Yes)	-0.031	0.012*	-0.056	<<0.001***	-0.024	0.228
Exercise practice (Yes)	0.099	<<0.001***	0.088	<<0.001***	0.046	0.007**

4.1.2 Model Diagnosis

Residual Plots

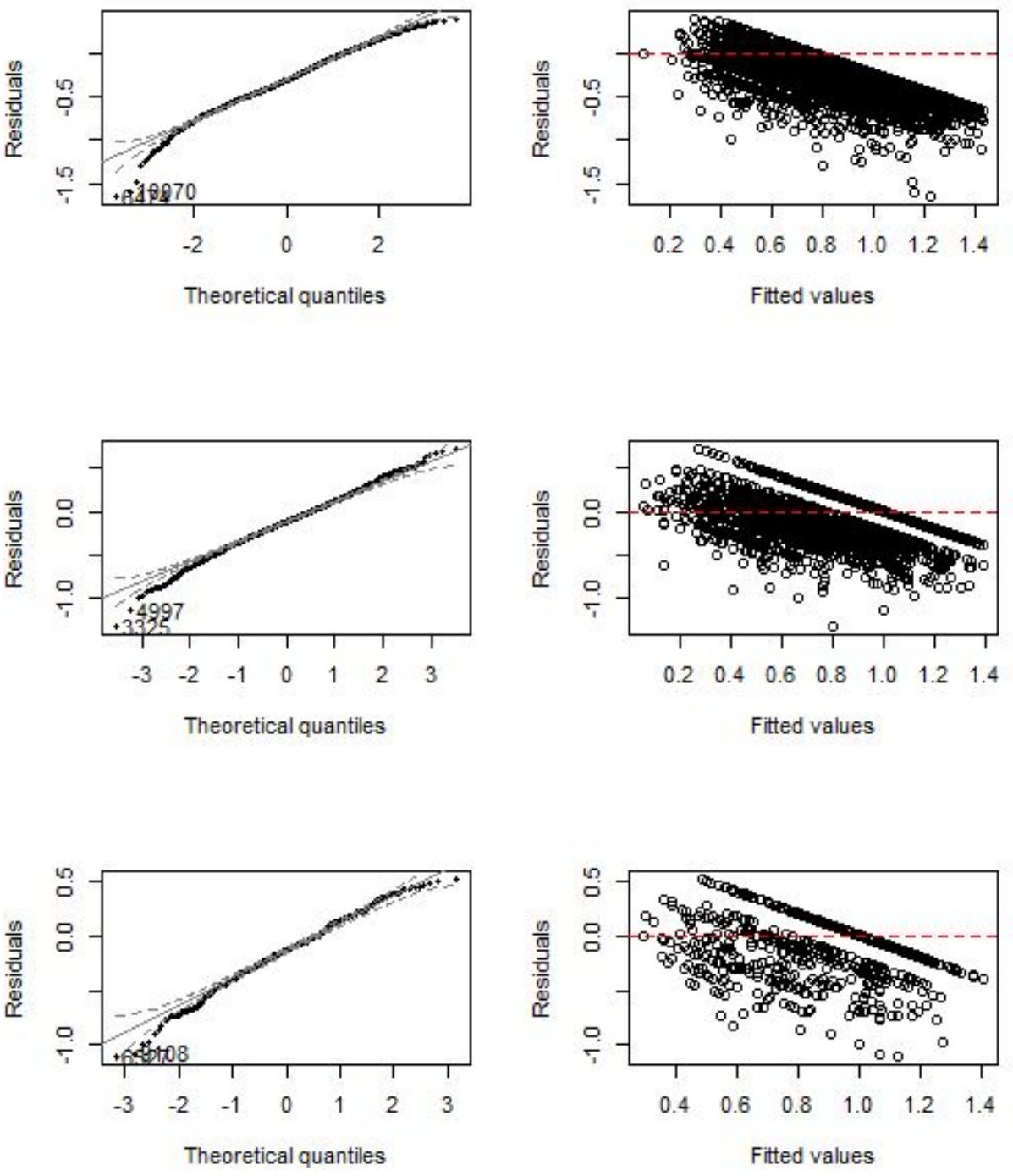

Fig. 4.2: Residuals plots of EQ-5D score for each wave, respectively.

To perform this analysis, censored observations were removed. As can be seen in the graphs in Figure 4.2, even though normality is not a requirement in these models, the qqplot shows a regular aspect for the residuals. However, the same is not true for the fitted values vs residuals plot, as the residuals appear to form a pattern.

Response variable analysis

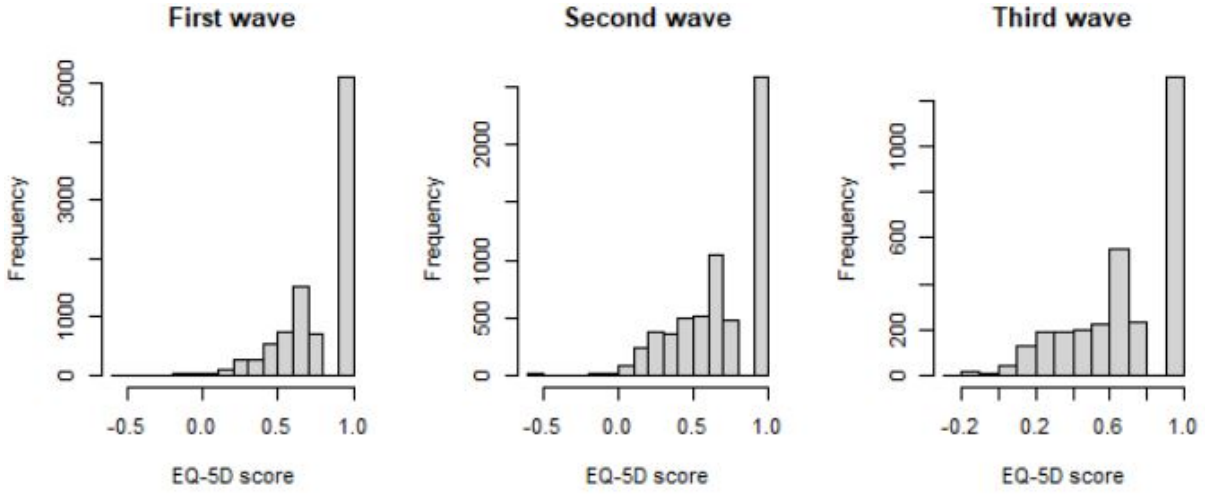

Fig. 4.3: Histogram of EQ-5D score for each wave, respectively.

The histograms show that the variable under study does not seem to be continuous either. Besides, the variable presents more mass points beyond the value 1. This fact may explain the appearance of the residuals in the fitted values vs residuals plot.

Influential Observations

Through the Dfbetas graphs mentioned in the appendice B it is noticed that there are no influential observations for the first wave data. Since, for the first wave there are 9297 observations and therefore the thereshold would be ± 0.21. For the other two waves, the same is not true. In the second wave the number of observation is 6196 and therefore the thereshold is ± 0.0250 and in the third wave the number of observations is 3078 and the thereshold is ± 0.036. For these two waves there are lots of influential observations and it does not make sense to remove them all from the model.

4.1.3 Variable interpretation

Tobit regression coefficients are interpreted similarly to OLS regression coefficients. However, the linear effect is on the uncensored latent variable, not the observed outcome. It is concluded that (the whole analysis assumes that all other variables are fixed):

- **Sex**
  The gender variable is significant for quality of life. The expected value of quality of life decreases in women when compared to men, for each wave. Therefore, quality of life is worse in females.

- **Age**
  The expected value of quality of life decreases for each additional year.

- **NUTS II**
  The expected value of quality of life decreases in the Açores, when compared to Lisboa, in the first wave. Whereas in the second wave, the expected value of the quality of life decreases in all the NUTS II when compared to Lisboa. And in the third wave it decreases in all except the Algarve, still compared to Lisboa.

- **Employment Status**
  In the first wave the expected value for quality of life decreases in the ”not employed” category and in the ”retired” category when compared with the ”employed” category. Whereas in the second and third waves, the expected value for quality of life only decreases in the ”not employed” category when comparing with the ”employed” category.
- **Years of education**
  The expected value of quality of life increases for each additional year of education.
- **Diseases**
  In the first and second wave, the expected value of quality of life decreases with the presence of the disease. In the third wave, there is one exception, in which the value of quality of life increases with the presence of the diseases blood pressure.
- **BMI**
  The expected value of quality of life decreases for each additional point of BMI.
- **Lifestyle habits**
  The expected value of quality of life increases with alcohol intake and practice exercise. The opposite happens with smoking habit.

4.1.4 Modelling score quality of life in a longitudinal approach

In order to graphically analyse the profile of the individuals regarding the score of quality of life (figure 9) , 20 individuals that participate in the three waves were randomly selected. Observing the graphic, only 16 lines can be distinguished, which is due to the fact that in these 20 individuals, 5 have a profile in which all the values of the score in the three waves are equal to 1. Furthermore, in most cases it is possible to see that the value of the score does not vary much from wave to wave.

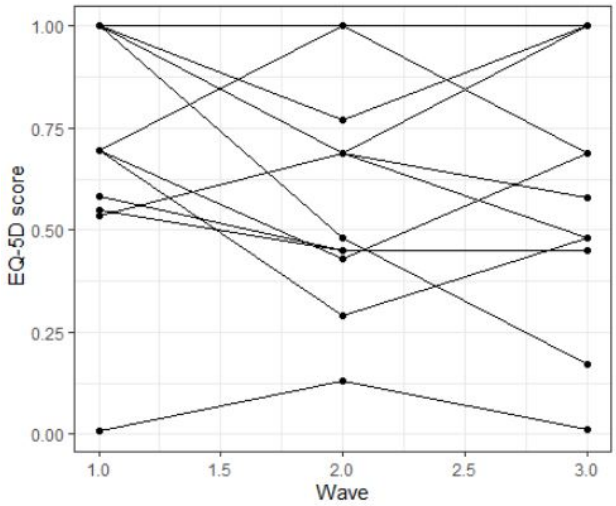

Fig. 4.4: Line graph of the profile of 20 individuals for the score of quality of life.

All analyses were prepared with the R software, except for the estimation of the quality of life model in the longitudinal approach. This analysis was performed in the Stata software, since no options were found to elaborate the longitudinal tobit model in the R software.

Table 4.5: Estimation of the quality of life score model.

Variable	$\hat{\beta}$	p-value
Intercept	1.433	<<0.001***
Sex Female	-0.127	<<0.001***
Age	-0.004	<<0.001***
NUTS II Norte	-0.011	0.302
NUTS II Centro	-0.030	0.008 **
NUTS II Alentejo	0.013	0.446
NUTS II Algarve	0.002	0.926
NUTS II Açores	-0.046	0.001***
NUTS II Madeira	-0.030	0.027*
Employment Status Not employed	-0.032	0.002 **
Employment Status Retired	0.017	0.093
Years of education	0.013	<<0.001***
Diabetes (Yes)	-0.024	0.028*
High blood pressure (Yes)	-0.019	0.023*
Cholesterol (Yes)	-0.020	0.009 **
Cardiac (Yes)	-0.077	<<0.001***
Mental (Yes)	-0.123	<<0.001***
Gastrointestinal (Yes)	-0.081	<<0.001***
Rheumatic (Yes)	-0.221	<<0.001***
Allergies (Yes)	-0.021	0.014*
Pulmonar (Yes)	-0.683	<<0.001***
BMI	-0.007	<<0.001***
Alcool intake (Yes)	0.046	<<0.001***
Smoking habit (Yes)	-0.070	<<0.001***
Exercise practice (Yes)	0.063	<<0.001***

4.1.5 Model Diagnosis

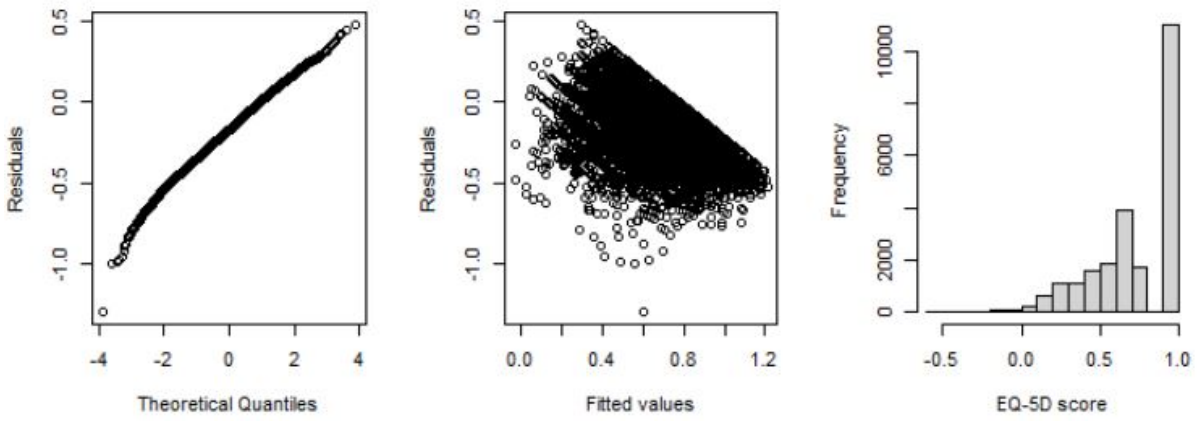

Fig. 4.5: Residuals plots and histogram of EQ-5D score.

As with cross-sectional approaches, the observations censored were removed for this analysis. In the

longitudinal approach, residuals continue to behave as expected with respect to the normality plot. While in the fitted values vs residuals plot a certain pattern is still found. Through the histogram, drawing the same conclusions as in the transversal approach, the response variable does not appear to be continuous and presents other mass points beyond the censure point. However, the coefficients associated to each variable were interpreted to understand if they are in accordance with what was expected.

Variable interpretation

It is concluded that (the whole analysis assumes that all other variables are fixed):

- **Sex**
  The expected value of quality of life decreases by 0.127 points in women when compared to men along the three waves.
- **Age**
  The expected value of quality of life decreases 0.004 points for each additional year. The variable age is significant for quality of life.
- **NUTS II**
  The quality of life score decreases in the Norte (0.011), Centro (0.030), Açores (0.046) e na Madeira (0.030) when compared to Lisboa. On the contrary, quality of life increases in the Alentejo e Algarve when compared to Lisboa.
- **Employment Status**
  The expected value of the quality of life score decreases for the unemployed (0.032) when compared with the employed. This same value increases for retired individuals (0.017), but this difference is not significant.
- **Years of education**
  The quality of life score increases for each additional year of education (0.013).
- **Diseases**
  The quality of life score decreases with the presence of all diseases. The quality of life score decreased 0.024 points in diabetics compared to non-diabetics.
- **BMI**
  In relation to BMI, for each point that it increases, the quality of life score decreases by 0.007 points.
- **Lifestyle habits**
  Regarding alcohol intake, the quality of life score increased by 0.046 points in individuals who drink alcohol compared to individuals who do not drink alcohol. Related smoking habit, there is a reduction of 0.070 points in the score of quality of life in individuals who smoke compared with individuals who do not smoke. Finally, the score of quality of life increased by 0.063 points in individuals who exercise in comparison with individuals who do not exercise.

4.2 Hospitalizations - Logistic Model

In order to obtain information about the hospitalizations the participants had been subject to, the following questions were asked: In the 1^{st} wave "Have you been hospitalized in the last 12 months?", in the 2^{nd} and 3^{rd} wave the question asked was "Have you been hospitalized since the last contact?".

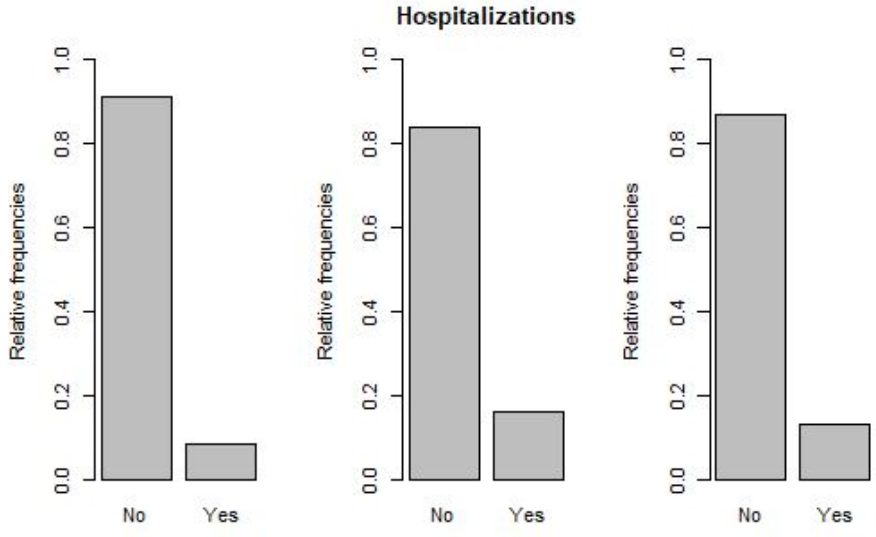

Fig. 4.6: Relative frequencies of hospitalizations corresponding to 1^{st}, 2^{nd} and 3^{rd} waves.

As expected, there were few individuals hospitalized. In the first wave 10655 individuals answered the question, 9732 were not hospitalized and 923 were. Therefore, the hospitalization proportion in the first wave was 8.7%. In the second wave, the hospitalization proportion was 16%, since 7547 answered and 1208 answered that they were hospitalized. Regarding the third wave, 5622 participants answered to the question and 737 of these were hospitalized, translating into a proportion of 13.1%.

Table 4.6: Weighted prevalences of hospitalizations referring to diabetics and non-diabetics.

	Diabetics	Non-diabetics
Hospitalized	14.5%	7.76%
Non-hospitalized	85.5%	92.2%

Taking into account diabetic individuals, the proportion of non-hospitalized individuals is higher than the proportion of hospitalized individuals, but the same is true for non-diabetic individuals. This was an expected result, since the overwhelming majority of individuals were not hospitalized.

A logistic regression model of a longitudinal nature could be developed, with the response variable being ”Was or was not hospitalized”, this is a possible approach since the response variable is binary and by the existence of observations collected for the same individual over time. However, convergence problems were encountered with this model and so it was decided to use GEE models. As the interviews were not all conducted at the same spacing, it was deemed necessary to add a parameter to control for this - the exposure time variable. The exposure time variable is represented by 12 months at the first time point, followed by the difference between the dates of the second and first interviews and then the difference between the dates of the third and second interviews.

4.2.1 Model Estimation

When estimating the model of hospitalizations, it is believed that the dependency structure that should be applied is Autoregressive, since there is an order imposed by the three waves and this temporal order matters. However, this structure did not work in the data as the data was not complete because many individuals did not have a response in all waves. Alternatively a fixed or unstructured structure can

be used. To use the fixed structure it is necessary to calculate the residuals through the GLM model and to calculate their correlation matrix, finally the calculated structure is imposed to the GEE model.

The working matrix for the fixed structure was $\begin{bmatrix} 1 & 0.191 & 0.062 \\ 0.191 & 1 & 0.225 \\ 0.062 & 0.225 & 1 \end{bmatrix}$.

Alternatively, using the correlation matrix with unstructured structure only requires estimating the model with this imposition.

The working matrix for the unstructured structure was $\begin{bmatrix} 1 & 0.187 & 0.021 \\ 0.187 & 1 & 0.108 \\ 0.021 & 0.108 & 1 \end{bmatrix}$.

It is possible to observe that both the fixed and the unstructured structure are similar and are very close in nature to an autoregressive. Which is in line with what was initially thought. In the estimation of this model, the matrix with fixed structure was used.

Table 4.7: Estimation of the model of hospitalizations.

Variable	$\hat{\beta}$	$\widehat{OR}$	Robust Z	p-value
Intercept	-3.568	0.028	-17.364	<<0.001***
Sex Female	-0.054	0.947	-0.901	0.367
Age	0.004	1.004	1.468	0.142
NUTS II Norte	0.122	1.130	1.657	0.098*
NUTS II Centro	0.050	1.051	0.647	0.517
NUTS II Alentejo	-0.127	0.881	-1.120	0.263
NUTS II Algarve	-0.129	0.879	-0.822	0.411
NUTS II Açores	-0.060	0.942	-0.611	0.541
NUTS II Madeira	-0.466	0.628	-4.065	<<0.001***
Employment Status Not employed	0.095	1.100	1.155	0.248
Employment Status Retired	0.180	1.197	2.372	0.018**
Diabetes (Yes)	0.262	1.300	3.588	<<0.001***
High blood pressure (Yes)	0.137	1.147	2.207	0.027*
Cardiac (Yes)	0.523	1.687	7.989	<<0.001***
Mental (Yes)	0.182	1.200	2.772	0.006**
Gastrointestinal (Yes)	0.135	1.145	2.110	0.035*
Rheumatic (Yes)	0.294	1.342	4.762	<<0.001***
Pulmonar (Yes)	0.414	1.513	4.702	<<0.001***
BMI	0.012	1.012	2.318	0.020*
Alcool intake (Yes)	-0.346	0.708	-6.465	<<0.001***
Smoking habit (Yes)	0.262	1.300	4.530	<<0.001***
Time of exposure	0.041	1.042	13.501	<<0.001***

4.2.2 Model Diagnosis

Goodness-of-fit

The Hosmer-Lemeshow goodness-of-fit test was used. This test is used to assess whether the number of expected events from the logistic regression model reflects the number of observed events in the data. With a chi-square test statistic of 8.3147 and a p-value of 0.4034. This test revealed that the model fits the data well.

GLM Vs. GEE Model

Since the diagnosis of GEE models has not yet been explored much, there is not much information on this part of the analysis. That said, and given that the residuals of the GLM model and the GEE model, for the same set of data and variables, are virtually equal. The residuals are analysed using the GLM model, since no tools were found to analyse the residuals through the GEE model. Therefore the tools available through the GLM model were used.

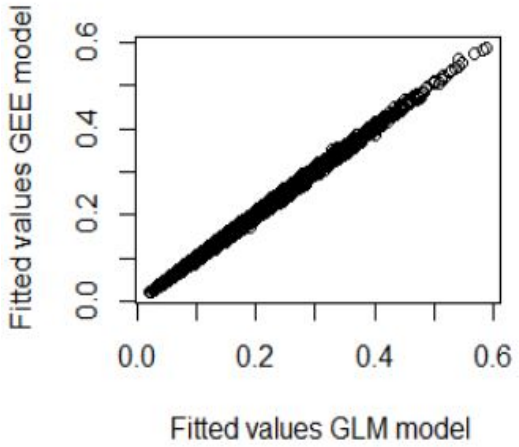

Table 4.8: Some fitted Values.

GLM model	GEE model
0.10198067	0.10009662
0.11836178	0.11665211
0.16971385	0.16502331
0.05147464	0.05128555
0.08933098	0.08875947
0.13757464	0.14204310

Fig. 4.7: Fitted Values GLM vs GEE model.

Since the fitted values of both models were very similar, proceeded to analyse the GEE residuals using GLM model tools.

Residual plots

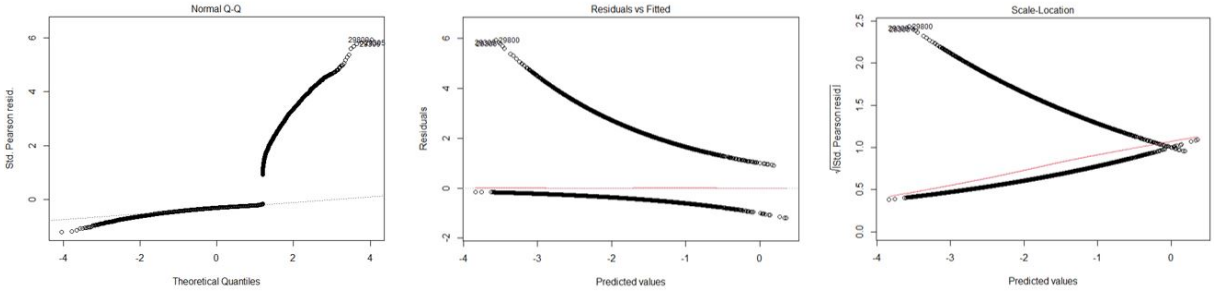

Fig. 4.8: Residuals plots

Residuals analysis is not an easy analysis to deal with in logistic regression models. Therefore, when the assumptions are met and the goodness-of-fit test revealed that the model is adequate, there is little reason to have residuals that are unreasonable.

Linearity

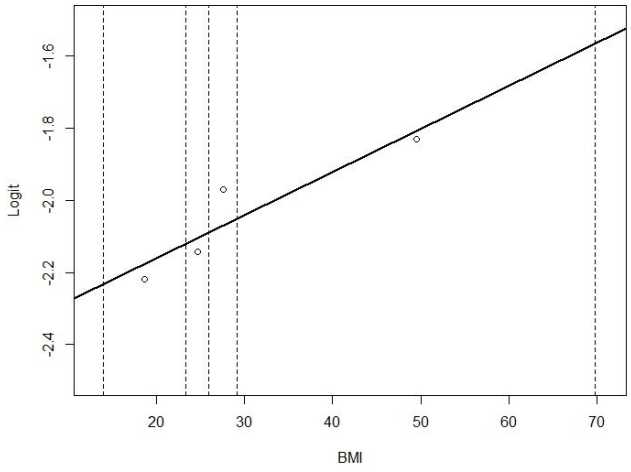

Fig. 4.9: Linearity of logit

A graph was drawn up for the linearity of the logit for the BMI variable (continuous quantitative variable).
The logit was calculated as follows: $Logit(\frac{h=1}{h=0})$. Being h=1 the proportion of individuals who were hospitalized and h=0 those who were not, in each of the intervals of the BMI variable.
The intervals chosen were: $[Min, Q_1], [Q_1, Q_2], [Q_2, Q_3], [Q_3, Max]$. The values of each point on the x axis is the midpoint of each of the intervals.

It can be seen from this graph that the linearity assumption of the logit is fulfilled.

Influential Observations

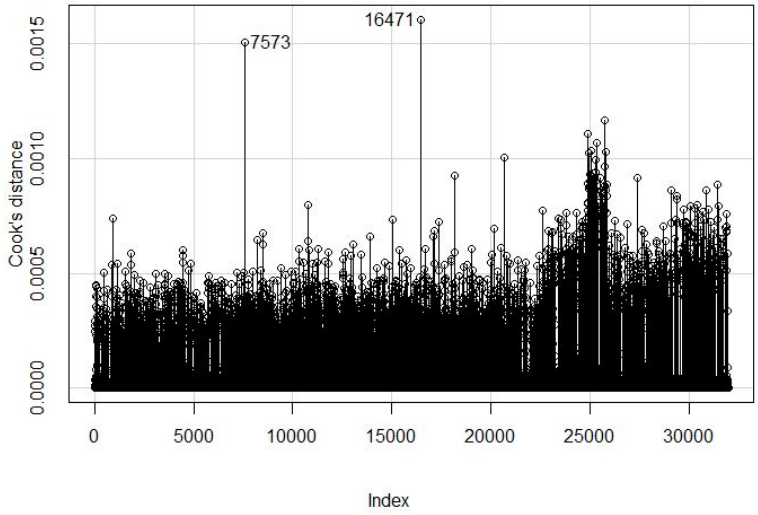

Fig. 4.10: Cook's Distance

Since all observations have Cook's distance values less than 1, there are no candidate observations for influential values.

4.2.3 Variables interpretation

It is concluded that (the whole analysis assumes that all other variables are fixed):

- **Sex**
  $\widehat{OR}$=0.947, which means that there is a 5.3% reduction in the chance of being hospitalized when comparing females with males.

- **Age**
  With a ten years increase in the age variable, $e^{10\times0.004} = 1.041$, there is a 4.1% increase in the chance of being hospitalized when compared to an individual ten years younger.

- **NUTS II**
  There is an increase in the chances of being hospitalized when comparing the Norte and Centro with Lisboa, of 13% and 5.1% respectively. For the others, there is a decrease in the chance of being hospitalized of 11.9%, 12.1%, 5.8% and 37.2% when comparing the Alentejo, Algarve, Açores and Madeira with Lisboa, respectively.

- **Employment Status**
  It can be seen that there is an increase in the chance of being hospitalized of 10% and 19.7% when comparing the not employed and the retired with the employed, respectively.

- **Diseases**
  The chance of being hospitalized increases whenever comparing an individual with the disease to an individual without the disease. Therefore, the chance of an individual being hospitalized increases by 30% for an individual with diabetes compared to an individual without diabetes. It increases by 14.7% for individuals with high blood pressure when compared to individuals who do not have the disease. The chance of being hospitalized increases by 68.7% for individuals with cardiac disease compared to individuals without cardiac disease. For individuals with mental disease the chance of being hospitalized increases by 20% compared to individuals without mental disease. The odds increase by 14.5% for those with gastrointestinal disease compared to those without. For individuals with rheumatic disease, the chance of being hospitalized increases by 34.2% compared with individuals without rheumatic disease. And finally, the chance of being hospitalized increases by 51.3% in individuals with pulmonar disease compared with individuals without this disease.

- **Body mass index**
  The chance of being hospitalized increases by 1.2% with a one point increase in BMI.

- **Alchool intake**
  Since the $\widehat{OR}$=0.708, it can be said that there is a 29.2% decrease in the chance of being hospitalized in individuals who consume alcohol compared with individuals who do not.

- **Smoking habit**
  There is a 30% increase in the chances of being hospitalized in individuals who smoke compared to those who do not smoke.

- **Time of exposure**
  The chance of being hospitalized increases by 4.2% for each extra unit of time.

4.3 Number of medical appointments - Poisson Model

In order to assess the impact of diabetes on healthcare utilisation, a model was developed taking into account the number of medical appointments attended by the patient. In this case, all medical appointments were considered, both those in the public and private sectors.

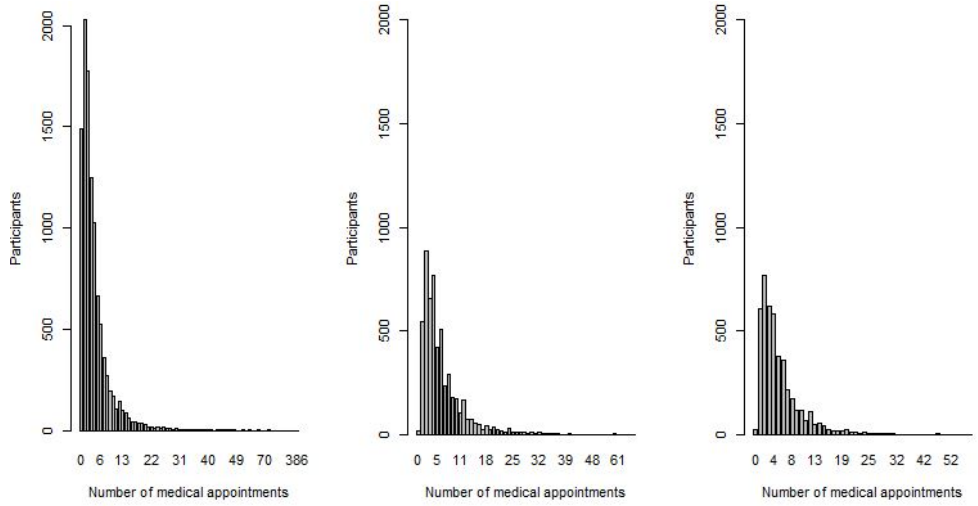

Fig. 4.11: Number of medical appointments corresponding to 1^{st}, 2^{nd} and 3^{rd} waves.

In the graphs in Figure 4.11, the overview of medical appointments attended by participants is depicted. When the questionnaire was applied, in the first wave the question asked was "How many medical appointments have you attended in the last 12 months?" and in the second and third waves the question was "How many medical appointments have you attended since the last contact?". For this reason, it was necessary to introduce in the Poisson regression model a term that regulates the fact that different periods are being considered. The offset of the logarithm of the exposure time was then used. This exposure time is expressed in months. So, at first wave the follow-up time was 12 months, then, in second wave, it was the difference between the date of the second interview and the first interview and at a third time it was the difference between the date of the third interview and the second interview.

Table 4.9: Number of medical appointments(NrM) per individuals(I).

NrM	I	NrM	I	NrM	I	NrM	I	NrM	I	NrM	I
0	660	13	122	26	17	39	4	54	1	150	1
1	1574	14	129	27	14	40	2	55	1	241	1
2	1943	15	92	28	9	41	2	57	1	355	1
3	1409	16	77	29	8	42	1	59	1	365	1
4	1446	17	53	30	15	43	3	60	1	367	1
5	838	18	67	31	8	44	2	61	2	386	1
6	853	19	40	32	12	45	2	63	1		
7	491	20	56	33	9	46	1	64	2		
8	470	21	34	34	8	48	4	66	1		
9	298	22	28	35	7	49	3	84	2		
10	259	23	18	36	6	50	1	102	1		
11	175	24	36	37	3	51	5	108	1		
12	257	25	16	38	2	53	1	110	1		

The Table 4.9 contains the number of medical appointments that each individual had from the beginning to the end of the study, and only individuals with complete data for the variables analysed are

considered.

Since there were four observations that were considered influential, these were removed. Still, the model had a tail with a very large weight on the left, the following table was analysed.

Table 4.10: Quantiles of the number of medical appointments.

Quantile	90%	91%	92%	93%	94%	95%	96%	97%	98%	99%
NrM	11	12	12	13	14	15	17	19	22.92	29

Data presented a heavy right tail and hence overdispersion that could not be handled with a Poisson model (Table 4.11). An adequate model to encompass such feature and still be able to describe adequately the left tail and central values of the response would require the presence of covariates that could explain the need of such a high number of medical appointments. The data collected did not seem to have such capacity and hence in an attempt to build a model that would fit well the large majority of the data, the 5% right tail of the data was not considered in the following model building.

4.3.1 Model Estimation

As in the previous section of the logistic model estimation, a correlation matrix with a fixed structure calculated using the GLM model was also used.

The working matrix used in this model was $\begin{bmatrix} 1 & 0.208 & 0.188 \\ 0.208 & 1 & 0.210 \\ 0.188 & 0.210 & 1 \end{bmatrix}$

Table 4.11: Estimation of the model of number of medical appointments.

Variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Robust Z	p-value
Intercept	-1.757	0.173	-48.378	<<0.001***
Sex Female	0.146	1.157	9.261	<<0.001***
Age	0.00006	1	0.095	0.924
NUTS II Norte	0.202	1.224	10.459	<<0.001***
NUTS II Centro	0.068	1.070	3.217	0.001 **
NUTS II Alentejo	0.035	1.036	1.168	0.243
NUTS II Algarve	0.006	1.006	0.1668	0.868
NUTS II Açores	-0.086	0.918	-3.219	0.001**
NUTS II Madeira	0.096	1.101	3.571	<<0.001***
Employment Status Not employed	0.082	1.085	3.657	<<0.001***
Employment Status Retired	0.133	1.142	6.581	<<0.001***
Diabetes (Yes)	0.1791	1.196	9.246	<<0.001***
High blood pressure (Yes)	0.122	1.130	7.719	<<0.001***
Cardiac (Yes)	0.191	1.210	10.313	<<0.001***
Mental (Yes)	0.204	1.226	11.516	<<0.001***
Gastrointestinal (Yes)	0.081	1.084	4.707	<<0.001***
Rheumatic (Yes)	0.163	1.177	9.811	<<0.001***
Pulmonar (Yes)	0.127	1.135	4.916	<<0.001***
Alcool intake (Yes)	-0.068	0.934	-4.824	<<0.001***
Exercise practice (Yes)	-0.063	0.939	-4.809	<<0.001***

4.3.2 Model Diagnosis

GLM Vs. GEE Model

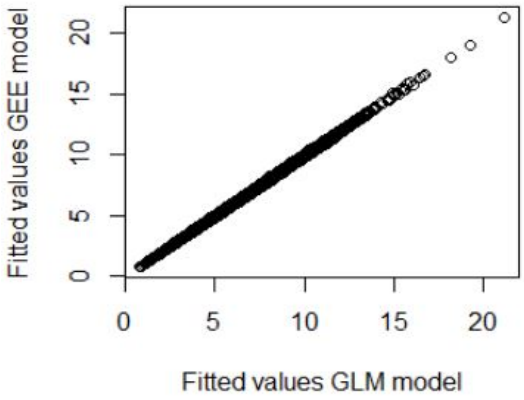

Fig. 4.12: Fitted Values GLM vs GEE model.

Table 4.12: Some fitted Values.

GLM model	GEE model
3.4989055	3.462233
4.811837	4.762044
5.559508	5.492854
2.773717	2.747872
4.451845	4.367743
3.424103	3.400911

As mentioned in the previous sub-section, there were no tools available to analyse the residuals through the GEE model. Since in this case also the residuals of the GEE and GLM models are found to be practically equal, the residuals of the GEE model were analysed using the tools of the GLM model.

Goodness-of-fit

If two nested models are fitted it is possible to compare their deviances. To this end, the null model was fitted and it was considered as a simple model. The models are compared using ANOVA and the chi-square test. Since the p-value is $< 2.2e^{-16}$, it is concluded that there is a significant difference between the models. However, the deviances are quite close. The deviance of null model is 47354 and full model deviance is 40908, then the difference between them being 6446.4. That said, the full model fits the data better than the simple model, however the difference in deviance is not that great and it may not be worth putting so many variables into the model.

Normality

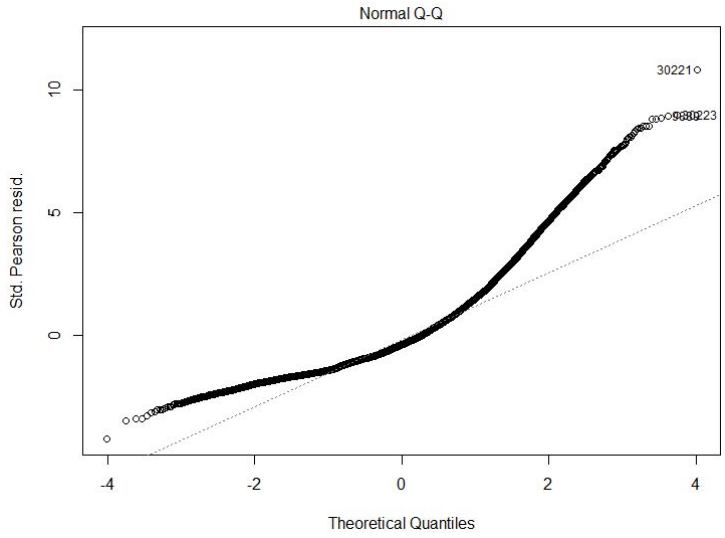

Fig. 4.13: Normality plot for the model of number of medical appointments.

The assumption of normality does not need to be fulfilled. Even so, the residuals should behave in a more regular way after being standardized, having a mean equal to 0 and have the same value for the

standard deviation. Graphically they should look more like an N(0,1) even though it is not necessary to follow this distribution. From the graph it can see that there are residuals with very high values, which reveals that there are observations that the model is not being able to follow.

Homoscedasticity

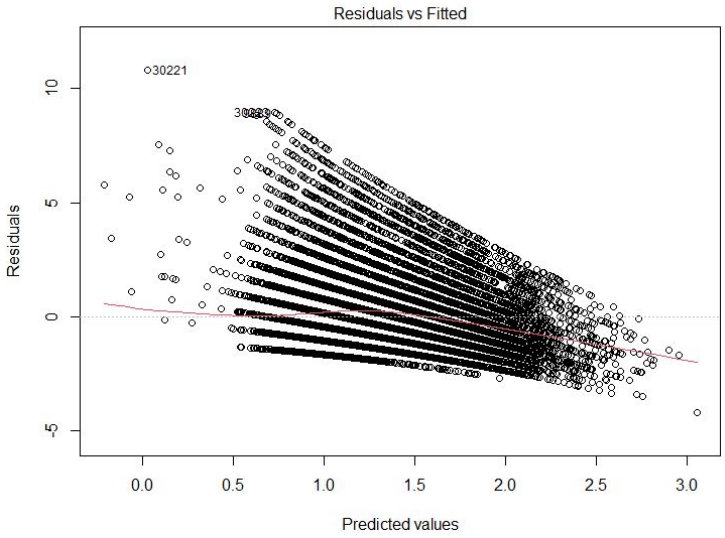

Fig. 4.14: Homoscedasticity plot for the model of number of medical appointments.

The homogeneity of variance does not need to be satisfied either. As mentioned above with respect to normality, in this case it can also be seen that the residuals do not behave in a very regular way.

Outliers

Through the outlier test with Bonferroni correction, obtained a test statistic of 2.682036 and a p-value $< 2.2e^{-16}$, one observation was excluded as it was considered an outlier.

Influential Observations

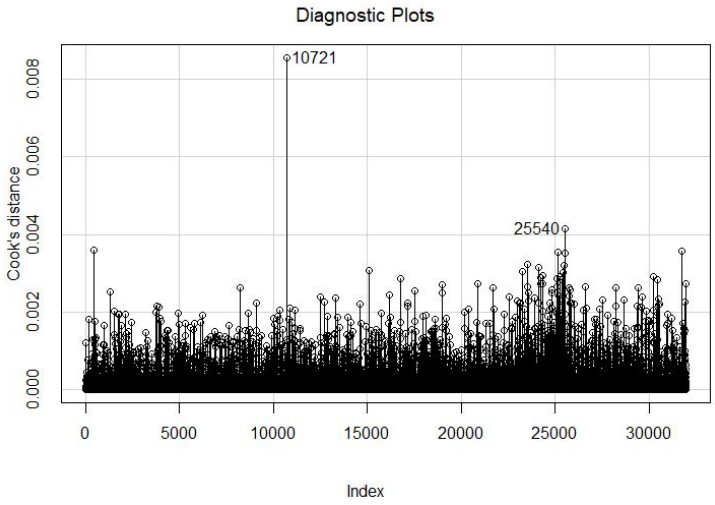

Fig. 4.15: Cook's distance for the model referring to number of medical appointments.

It was necessary to find out if there were any influential observations. Since all values are below 1, it was not necessary to remove any further observations.

The residuals does not behave in the most appropriate way. Although the model captures the type of information that the variables have about the response, the model does not have the flexibility to track the variability of the response. A model that accommodates this type of variability would be needed. A model with a quasi-poisson distribution was still developed, but the variability was not accommodated anyway.

4.3.3 Exploring the model variables

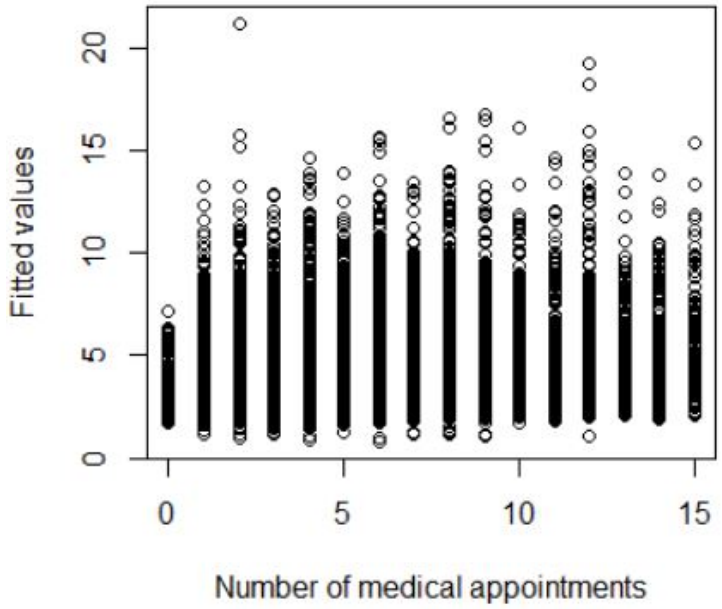

Fig. 4.16: Plot of fitted values vs observed values.

Table 4.13: Mean and variance of the number of medical appointments.

NrM	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Mean	2.60	3.16	3.69	3.93	4.33	4.35	4.84	4.67	5.16	5.14	5.04	5.09	5.59	4.93	5.24	5.50
Variance	0.45	1.77	2.59	2.88	3.61	3.62	4.55	4.33	6.09	6.02	5.03	6.43	7.71	4.57	6.25	6.68

The Figure 4.16 shows that the adjusted values increase very slightly compared to the target. As the Table 4.13 shows, the variance and the mean are practically constant, which means that the variables introduced to explain the number of medical appointments according to the Poisson model do not lead to an increase in the estimated number of medical appointments. The explanatory variables have little effect on the number of medical appointments.

The number of medical appointments was analysed in relation to each of the explanatory variables to explore the relationship between the variables.

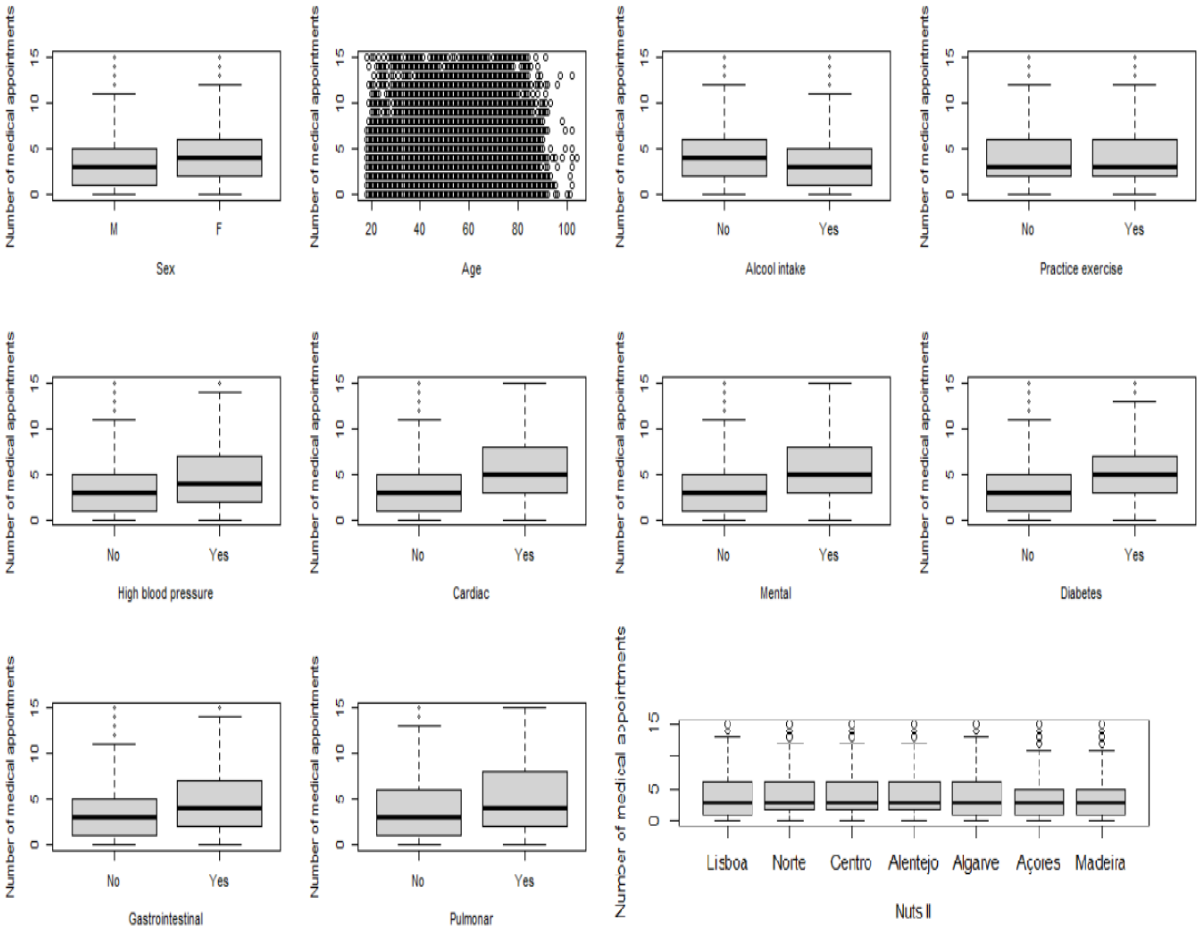

Fig. 4.17: Plot of number of medical appointments vs explanatory variables

These graphs show that the distribution of the number of medical appointments between gender, age, alcohol intake, physical exercise and the presence or absence of diseases is practically equal. Therefore, it is normal that the variables have little effect on the number of medical appointments varying.

4.3.4 Variables interpretation

Even if the residuals do not behave ideally, it is meaningless to interpret the estimates of parameters associated with each variable. Still, it is important to analyse whether the variables, through the model mentioned in the table 4.11, are behaving as expected. The expected number of medical appointments decreases in Açores compared with Lisboa, decreases with alcohol intake and with practice exercise. On the other hand, the expected number of medical appointments increases with age and the presence of diseases. It also increases in not employed and retired people compared to employers and increases in women compared to men.

5. Discussion and Conclusion

For this study, EpiDoC data, collected on a large scale, was used. This data becomes an advantage for this study, since the data is longitudinal and therefore one can assess the impact of diabetes on quality of life and health resources with medical follow-up over a few years. Although the data collection was done with the aim of studying rheumatic diseases, it is possible to analyse diabetes as there are variables on it.

As diabetes is a very harmful disease because it causes so many other complications, it is a subject that should be well studied and addressed to make people aware of the risks of this disease and its risk factors (Organization, 2000).

Therefore, the aim of this study was to characterise the diabetes population among Portuguese adults taking into account the geographical distribution, socio-demographic characteristics and comorbidities associated with the disease. Also, to assess the impact of diabetes on quality of life and the occurrence of hospital admissions and the number of medical appointments.

Regarding the prevalence of diabetes among adult Portuguese, it is concluded that gender is not a relevant factor for the presence of diabetes. Nevertheless, there is a higher proportion of women in both the diabetic (57.1%) and non-diabetic (52.3%) groups. Mean age is higher in diabetics (66.38 years). The most represented age group among diabetics is from 66 to 75 years (29.4%). The diabetics were more represented in the North (34.2%). Regarding Ethnicity, it was expected that there would be more caucasian people in diabetics (96.1%) group, since Portugal has in its constitution a prevalence of people of caucasian ethnicity. However, there were more caucasian people in non-diabetics group. Among diabetic group, the proportion of married individuals (63.2%) is higher than other marital status. Taking into account the years of education, the individuals who studied between 1 and 4 years are those with the highest proportion of diabetes (50.1%). However, it is possible that age is considered a confounding factor for education levels, since older people have a lower level of education. The highest prevalence of diabetics is found among the retired (65.4%), however age is a possible confounding factor, by the fact most diabetics are older, which is also the case for retired people. Regarding lifestyle habits, the highest proportion of diabetics is found among individuals who reported drinking alcohol (51.3%). The prevalence of diabetics who do not smoke (89.3%) is higher than those who smoke. And finally, the proportion of diabetics who do not practice exercise (77.7%) is also higher than that of those who do. The highest proportion of diabetics is found among those with a BMI level considered obese (41.5%). Taking into account the diseases, there is a higher proportion of diabetics in the presence of cholesterol (55.7%) and high blood pressure (65.6%). A result that was confirmed in the association rules, where it was concluded that among diabetics the most common situation is to have high blood pressure (66%) and cholesterol (59%). Regarding the distance matrix between diseases, it is concluded that the most evident groups of diseases were cardiac; diabetes; pulmonary; gastrointestinal; mental and cholesterol;

high blood pressure; rheumatic; and allergies.

For the study of quality of life, models were developed for each wave using the Tobit model. Problems were found with the residuals of these models, since the scores of the answers to the EQ-5D-3L questionnaire are not very distinct caused there to be more mass points beyond the point 1 that corresponded to the censoring point. These points make the plot of the residuals, even with the exclusion of the accumulation point at 1, manifest themselves in an intense way in the plot. Even so, the parameters associated with the explanatory variables were interpreted, which was the case. It was concluded that women have less quality of life than men. Quality of life gets worse for each year more, across all waves. In the first wave, the quality of life is worse in the Açores when compared to Lisboa. In the second wave, the quality of life score worsens in all regions when compared to Lisboa. In the third wave, the quality of life worsens in all the regions except for the Algarve. Regarding employment status, quality of life is worse for the not employed compared to the employed, and it is also worse for the retired, but only in the first and second waves. The quality of life score improves with each additional year of education. In the field of diseases, their presence reveals a worse quality of life. Except, in the third wave, where the score of quality of life improves in the presence of high blood pressure. The quality of life worsens in all waves in the presence of diabetes, even so, the difference between having or not having the disease is not significant for quality of life in the second and third waves. Relating to BMI, the quality of life worsens for each additional point of BMI. As for lifestyle habits, quality of life improves with the ingestion of alcohol and with physical exercise. On the contrary, it worsens with smoking.

For the same purpose, the analysis of the quality of life score was also performed, but this time in a longitudinal approach. In this approach people are followed throughout the time in which the study took place and therefore it is possible to see a better evolution of each parameter assessed.

As in the cross-sectional approach, it was also found that quality of life decreases in females, for each additional year, for each additional point of BMI and with smoking. On the other hand, and corroborating the transversal analysis, it increases for each additional year of education, with the ingestion of alcohol and with the practice of exercise. The longitudinal analysis shows that the score for quality of life decreases in the Norte, Centro and Açores and Madeira when compared with Lisboa. The quality of life increases in Alentejo and Algarve. Quality of life decreases with the presence of all diseases.

To assess the occurrence of hospitalizations and the number of medical appointments generalized linear models were used. When the project was being developed, limitations concerning the R program were experienced, as the linear mixed-effects models, one of the approaches that could be used, was ineffective as the models took too long to run which was unsustainable. This being said, the GEE approach was used to develop these models. The models were estimated using the gee package, but a disadvantage of this package is that it does not directly provide p-values but provides the test statistics, which can be used to find the p-values.

Regarding hospitalizations, it was concluded that there is a reduction (5.3%) in the chance of being hospitalized in women compared to men. For every 10 years plus, there is an increase (4.1%) in the chance of being hospitalized. The chance of being hospitalized increases in the Norte (13.4%) and in the Centro (5.1%) when compared with Lisboa. On the contrary, there is a reduction in the chance in Alentejo (11.9%), Algarve (12.1%), Açores (5.8%) and Madeira (37.2%) when compared with Lisboa. There is an increase in the chance of being hospitalized among the non-employed (10%) and the retired (19.7%) when compared to the employed. The presence of diseases increases the chance of being hos-

pitalized. It increases for diabetes (30%), high blood pressure (14.7%), cardiac disease (68.7%), mental disease (20%), gastrointestinal (14.5%), rheumatic (34.2%) and pulmonary disease (51.3%). The BMI increases (1.2%) the chance of being hospitalized, for each extra point. Regarding lifestyle habits, with alcohol intake there is a reduction (29.2%) in the chance of being hospitalized. While there is an increase (30%) in the chance of being hospitalized in those who smoke. Finally, with an increase of one unit of time, the chance of being hospitalized increases (4.2%).

To analyse the number of medical appointments, it was calculated the quantiles and removed the 5% right tail of the data, since data collected did not seem to have such capacity and hence in an attempt to build a model that would fit well the large majority of the data.

Therefore, only individuals who attended 15 or fewer medical appointments during the study were analysed. This model also revealed problems in terms of the residuals, since the residuals did not behave as expected, which can be explained by the small difference in the number of medical appointments between the categories of explanatory variables. Even so, it was analysed whether the variables were behaving as expected. The expected number of medical appointments increases with age, with the presence of diseases and in the non-employed and retired compared to the employed, which is as expected. It also increases in women compared to men, since in general women take more care of themselves, which was also expected. On the other hand, the expected number of medical appointments decreases with the ingestion of alcohol, which can be justified by the fact that if the individual had more problems the person might not drink alcohol and would go to more medical appointments. It also decreases with the practice of physical exercise and decreases in the Açores compared to Lisboa, all that was expected.

To conclude, this study was important, since diabetes has been increasing in the Portuguese population. Through this study, it was concluded that diabetes affects the quality of life, the number of medical appointments and the number of hospitalizations, in the presence of other variables.

Bibliography

Autumn, Statistical Inference (2016). "Lecture 27 — Poisson regression The Poisson log-linear model Statistical inference". In: pp. 1–4.

Bagley, Steven C. et al. (2001). "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain". In: *Journal of Clinical Epidemiology* 54.10, pp. 979–985.

Ballinger, Gary A. (2004). "Using Generalized Estimating Equations for Longitudinal Data Analysis". In: *Organizational Research Methods* 7.2, pp. 127–150.

Bock, Tim (n.d.). *What is a Distance Matrix?*

Brown, G. C. et al. (2000). "Quality of life associated with diabetes mellitus in an adult population". In: *PubMed*.

Brutsaert, Erika F. (2020). *Diabetes mellitus (DM)*.

Cabral, Maria Salomé (2019). *Apontamentos da Unidade Currícular Modelos Lineares e Extensões*.

Centers for Disease Control and Prevention [CDCP] (2014). "Smoking and Diabetes How Smoking Causes Type 2 Diabetes What Smoking Means To People With Diabetes". In: *Surgeon General's Report on Smoking and Health* 1-2.

Colberg, Sheri R. et al. (2016). "Physical activity/exercise and diabetes: A position statement of the American Diabetes Association". In: *Diabetes Care* 39.11, pp. 2065–2079.

Create, Title et al. (2022). "Package 'ggplot2'". In:

Diabetes, Observatório Nacional da (2016). *Diabetes Factos e Números - O ANO DE 2015*.

Dias, Sara Simões et al. (2018). "Cohort profile: The epidemiology of chronic diseases cohort (EPIDOC)". In: *International Journal of Epidemiology* 47.6, 1741–1742J.

EuroQol Research Foundation (2018). "User GuideEQ-5D-3L User Guide". In: pp. 1–33.

Fagerland, Morten W. and David W. Hosmer (2012). "A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistic regression models". In: *Stata Journal* 12.3, pp. 447–453.

Ferreira, Pedro Lopes et al. (2013). "Contributos para a validação da versão Portuguesa do EQ-5D". In: *Acta Medica Portuguesa* 26.6, pp. 664–675.

Fitzmaurice, Garrett M. et al. (2004). "Applied Longitudinal Analysis". In:

Fox, John (McMaster University) and Sanford (University of Minnesota) Weisberg (2011). "Chapter 6: Diagnosing Problems in Linear and Generalized". In: *An R Companion to Applied Regression*, pp. 285–328.

Functions, Description and Christian Kleiber (2022). *Package ' AER '*.

Garg, Anisha (2018). "Complete guide to Association Rules". In: *Towards Data Science*.

Generalized, Title and Estimation Equation (2022). "Package ' gee '". In: pp. 1–5.

Hardin, James W. and Joseph M. Hilbe (2002). *Generalized estimating equations*, pp. 1–222.

Hosmer, David W. and Stanley Lemeshow (2000). *Applied Logistic Regression - Hosmer, Lemeshow*.

John, Author et al. (2022). "Package 'car'". In:

Klein, John P. and Melvin L. Moeschberger (1997). "Censoring and Truncation". In: pp. 55–82.

Lumley, Author Thomas (2021). “Package ‘ survey ’”. In:

Lusiadas (2022). *Conheça os tipos de diabetes*.

Mendes Leal, Margarida (2019). *Apontamentos da Unidade Curricular Análise Exploratória de Dados Multivariados*.

Mosel, Stacy (2022). “Alcohol Diabetes: Can Alcohol Cause Diabetes?” In:

Notes, Detailed Lecture et al. (2015). “The Poisson Regression Model”. In: 7.3, pp. 1–19.

Organization, World Health (2000). *The challenge of diabetes*.

Pearson, Standardized and Standardized Pearson (n.d.). “Residuals in glm The terminology In the normal case”. In: (), pp. 2–5.

Raposo, João Filipe (2020). “Diabetes: Factos e Números 2016, 2017 e 2018”. In: *Revista Portuguesa de Diabetes* 15.1, pp. 19–27.

Santos, Joana et al. (2017). “Diabetes: Socioeconomic inequalities in the Portuguese population in 2014”. In: *Acta Medica Portuguesa* 30.7-8, pp. 561–567.

Smith, Douglas A. and Robert Brame (2003). “Tobit models in social science research: Some limitations and a more general alternative”. In: *Sociological Methods and Research* 31.3, pp. 364–388.

STATA (2013). “Postestimation commands”. In: pp. 1–5.

Stephenson, Maintainer Alec (2022). “Package ‘ evd ’”. In:

Sugawara, Etsuko and Hiroshi Nikaido (2014). “Properties of AdeABC and AdeIJK efflux systems of Acinetobacter baumannii compared with those of the AcrAB-TolC system of Escherichia coli”. In: *Antimicrobial Agents and Chemotherapy* 58.12, pp. 7250–7257. arXiv: `arXiv:1011.1669v3`.

Tunceli, Kaan et al. (2005). “The impact of diabetes on employment and work productivity”. In: *Diabetes Care* 28.11, pp. 2662–2667.

University, The Pennsylvania State (2018). *Applied Regression Analysis*.

Appendices

A Mixed effects models - Longitudinal Data

A.1 Logistic regression model

$$Y_{ij} \sim Bernoulli(p_{ij}), \quad i=1,...,n, \quad j=1,...,t$$

$$p_{ij} = Pr(y_{ij} = 1) = E(y_{ij}) \tag{1}$$

Introducing the random effect for the same reason mentioned in the poisson regression models (Fitzmaurice et al., 2004):

$$Y_{ij}|b_{0i} \sim Bernoulli(p_{ij})$$

$$p_{ij} = Pr(y_{ij} = 1|b_{0i}) = E(y_{ij}|b_{01}) \tag{2}$$

Therefore, the mixed-effects logistic model can be written as follows:

$$\text{logit}(Pr(y_{ij} = 1|b_{0i})) = \beta_0 + b_{0i} + \sum_{l=1}^{m}\sum_{i=1}^{n}\beta_l x_{li} + \sum_{r=1}^{s}\sum_{i=1}^{n}\sum_{j=1}^{t}\beta_{rj}x_{rij}, \tag{3}$$

where,
l=1,...,m, represents the fixed effect covariates, that do not depend on time.
r=1,...,s , represents the random effect covariates, that depend on time.
i=1,...,n, represents the individual
j=1,...,t, represents the data collection moments.
$b_{0i} \sim N(0, \sigma^2_{B0})$

A.2 Poisson regression model

$$y_{ij} \sim Poisson(\lambda_{ij}) = Poisson(e^{\mu_{ij}}), \quad i=1,...,n, \quad j=1,...,t$$

$$p(y_{ij}) = \frac{(e^{\mu_{ij}})^{y_{ij}}}{y_{ij}!}e^{-\mu_{ij}}, \tag{4}$$

where $\lambda_{ij} = e^{\mu_{ij}}$ is the expected number of a given event for individual i at time j.

Due to several observations for the same individual there is overdispersion and the variance exceeds the expected value (Fitzmaurice et al., 2004).

$$Var(y_{ij}) = \phi e^{\mu}_{ij} > e^{\mu}_{ij}, \phi > 1 \tag{5}$$

We deal with this situation by introducing random effects in the model, which will represent inter-individual variability.

So,

$$y_{ij}|b_{0i} \sim Poisson(e^{b_{0i}+\mu_{ij}})$$

$$p(y_{ij}|b_{0i}) = \frac{(e^{b_{0i}+\mu_{ij}})^{y_{ij}}}{y_{ij}!}e^{-(e^{b_{0i}+\mu_{ij}})} \tag{6}$$

Therefore, the mixed-effects poisson model can be written as follows:

$$\log E(y_{ij}|b_{0i}) = \beta_0 + b_{0i} + \sum_{l=1}^{m}\sum_{i=1}^{n}\beta_l x_{li} + \sum_{r=1}^{s}\sum_{i=1}^{n}\sum_{j=1}^{t}\beta_{rj}x_{rij} + \log(Tij), \tag{7}$$

where,

l=1,...,m, represents the fixed effect covariates, that do not depend on time.

r=1,...,s , represents the random effect covariates, that depend on time.

i=1,...,n, represents the individual

j=1,...,t, represents the data collection moments.

Log(T_{ij}) is defined as an offset. In order to control the exposure time.

$b_{0i} \sim N(0, \sigma^2_{B0})$

B Influential observations - Dfbetas

B.1 First wave

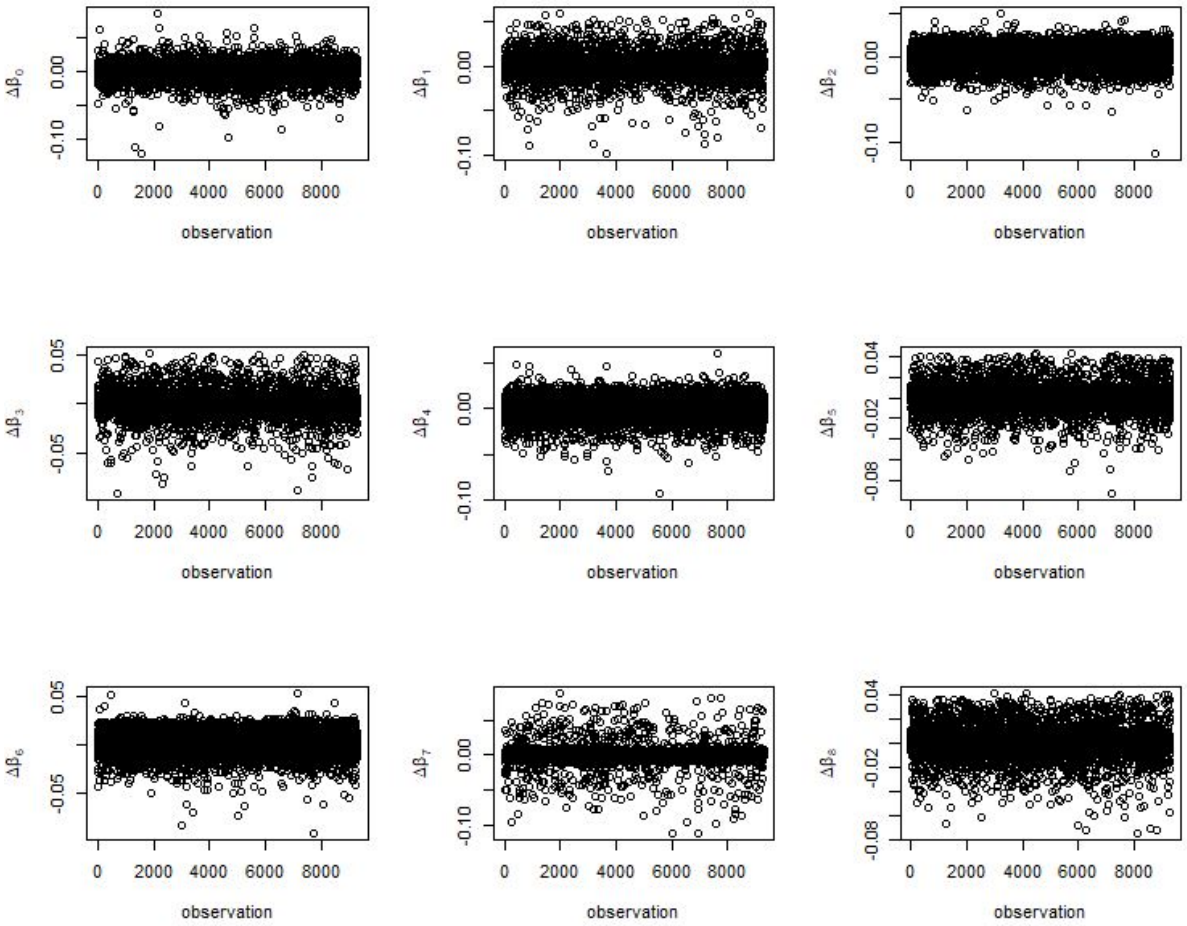

Fig. 1: Dfbetas for 1^{st} wave, $\beta = 0, ..., 8$.

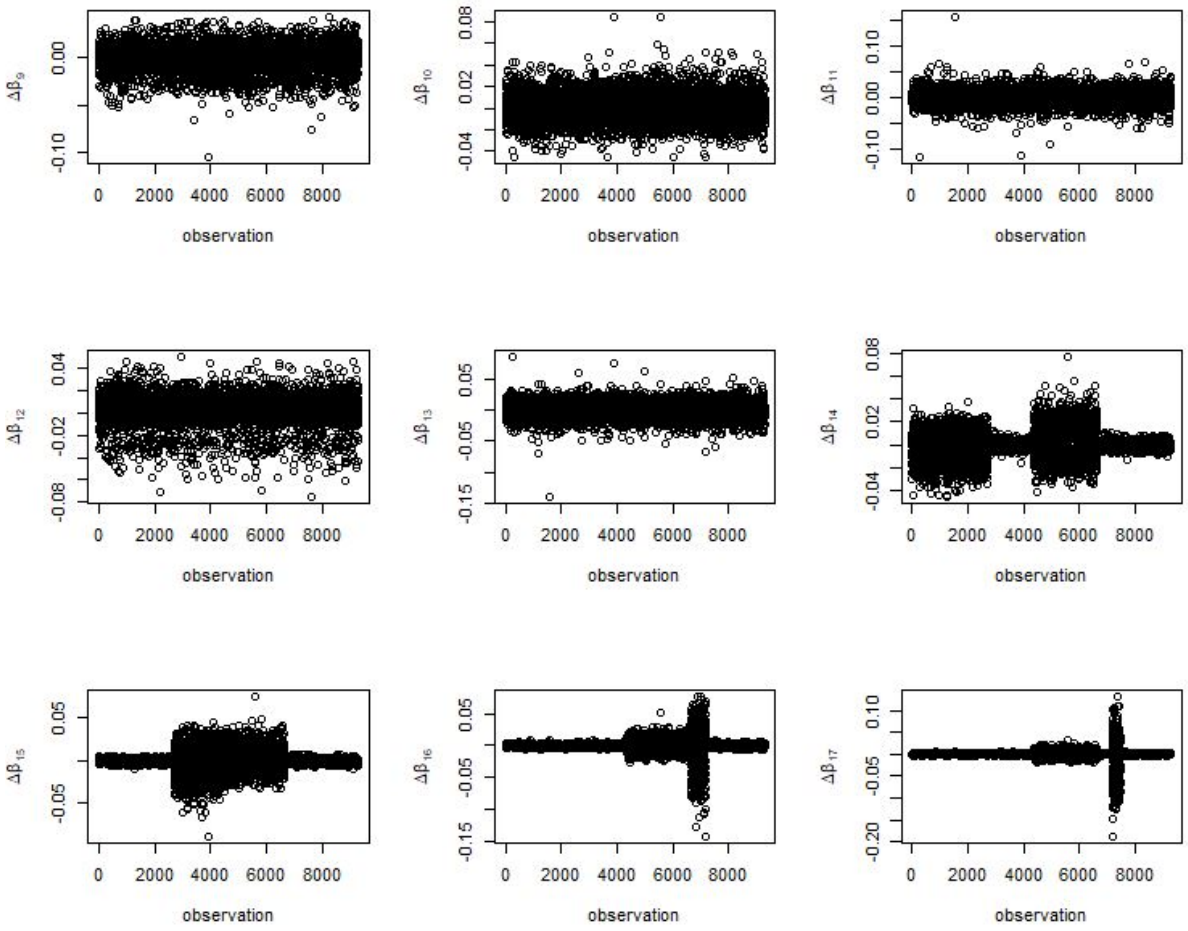

Fig. 2: Dfbetas for 1^{st} wave, $\beta = 9, ..., 17$.

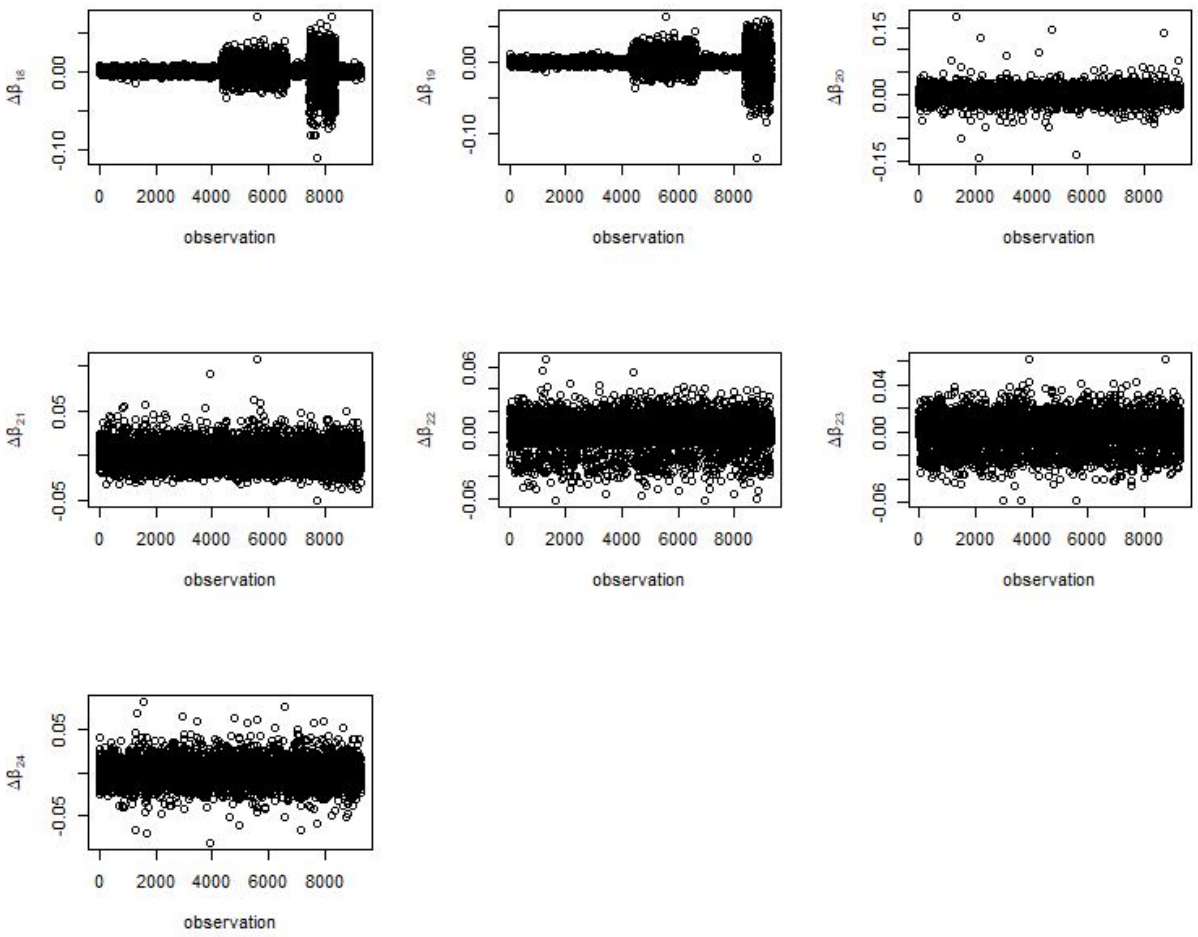

Fig. 3: Dfbetas for 1^{st} wave, $\beta = 18, ..., 24$.

B.2 Second wave

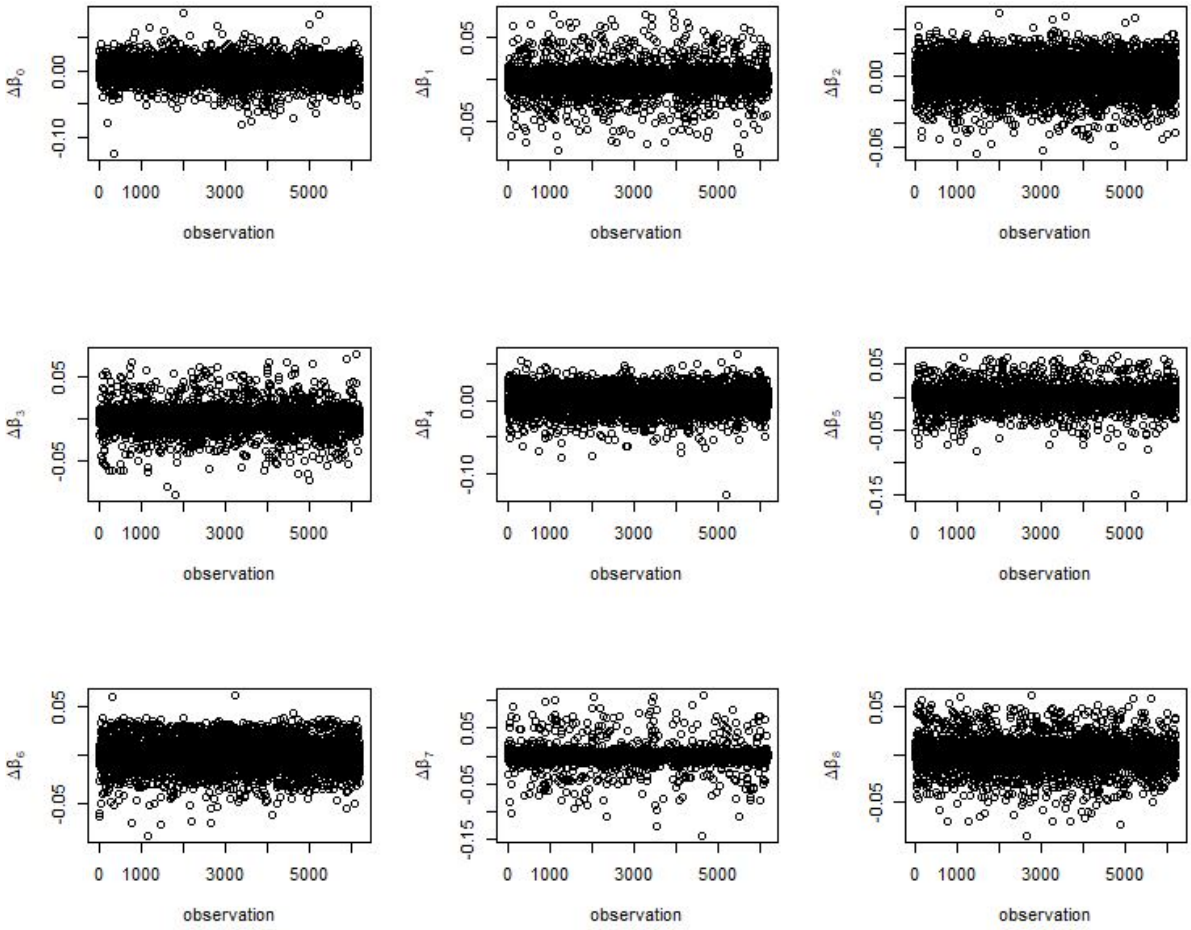

Fig. 4: Dfbetas for 2^{nd} wave, $\beta = 0, ..., 8$.

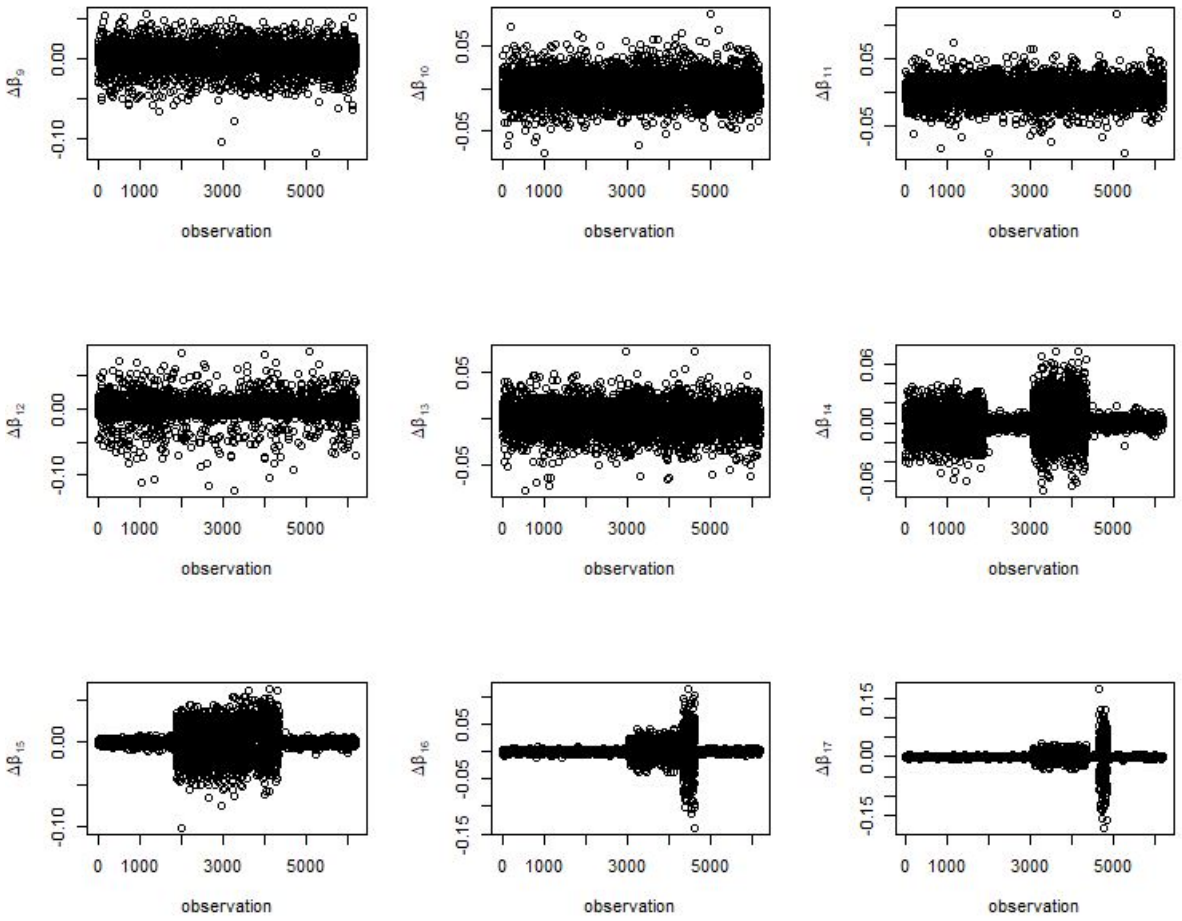

Fig. 5: Dfbetas for 2^{nd} wave, $\beta = 9, ..., 17$.

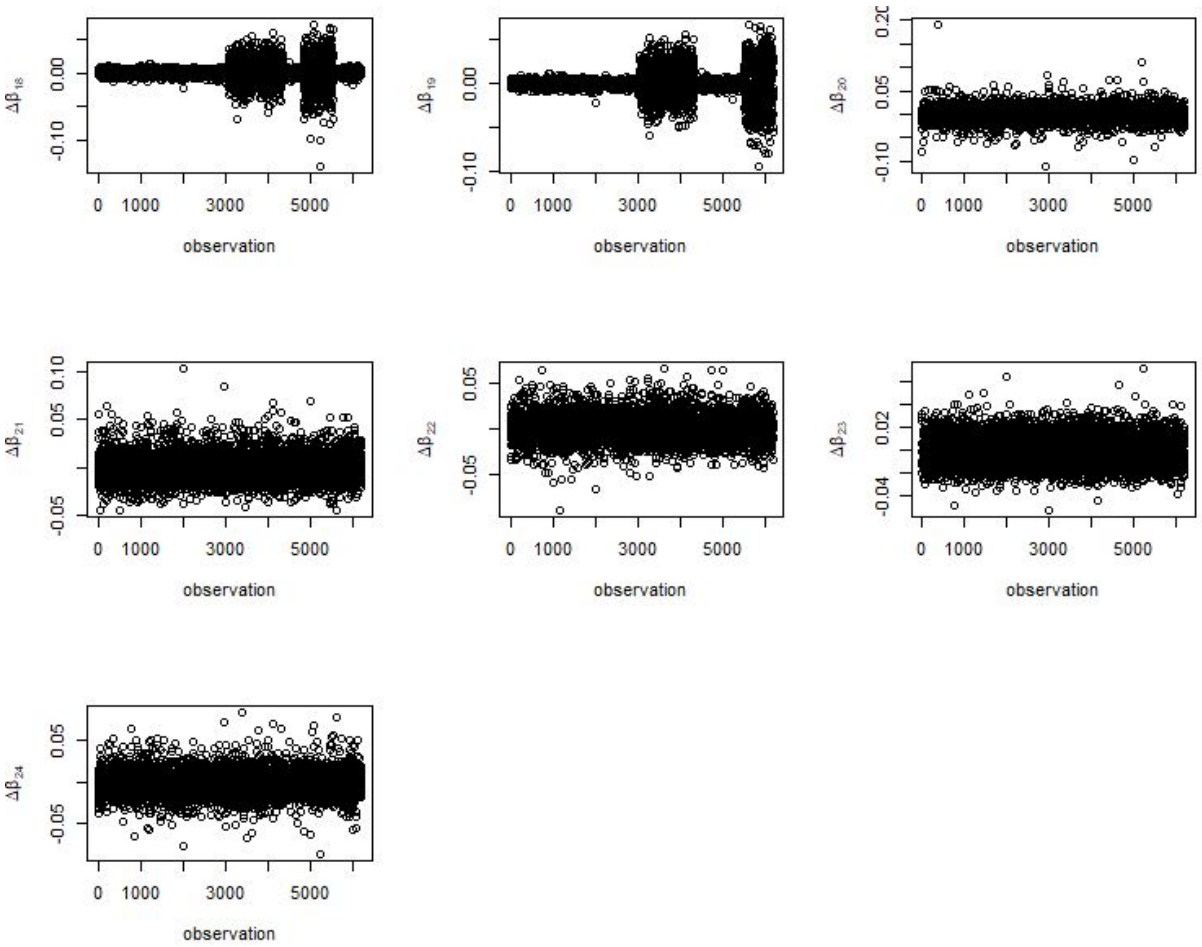

Fig. 6: Dfbetas for 2^{nd} wave, $\beta = 18, ..., 24$.

B.3 Thrid wave

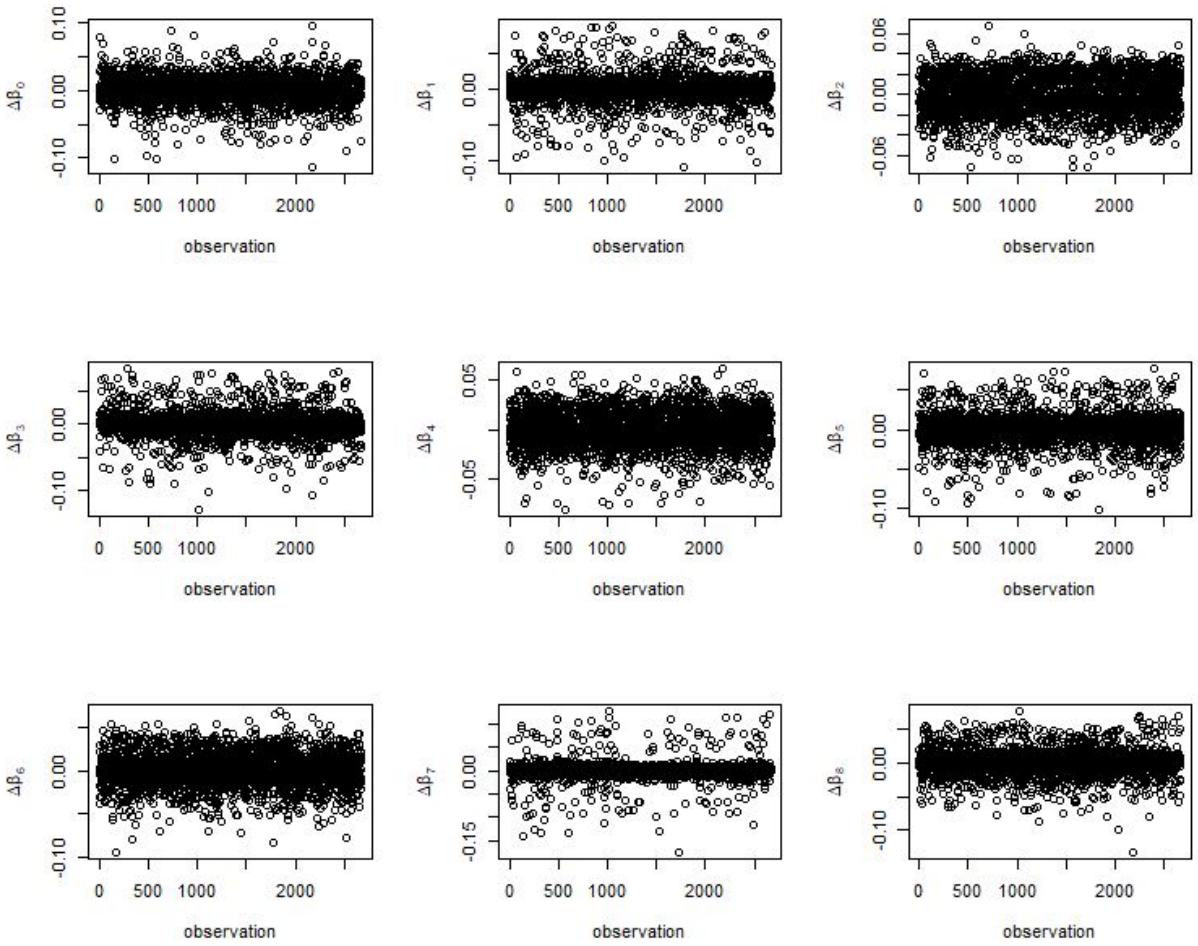

Fig. 7: Dfbetas for 3^{rd} wave, $\beta = 0, ..., 8$.

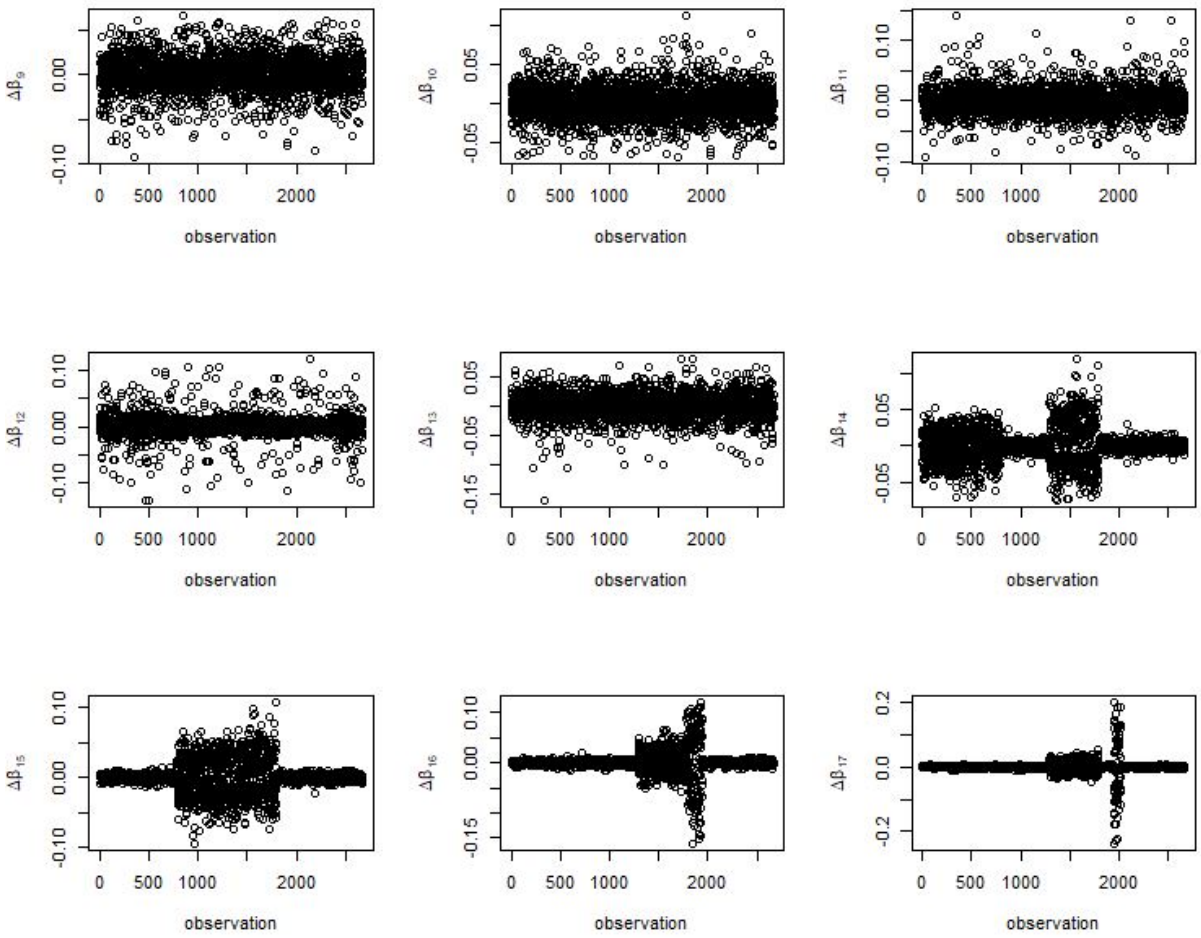

Fig. 8: Dfbetas for 3^{rd} wave, $\beta = 9, ..., 17$.

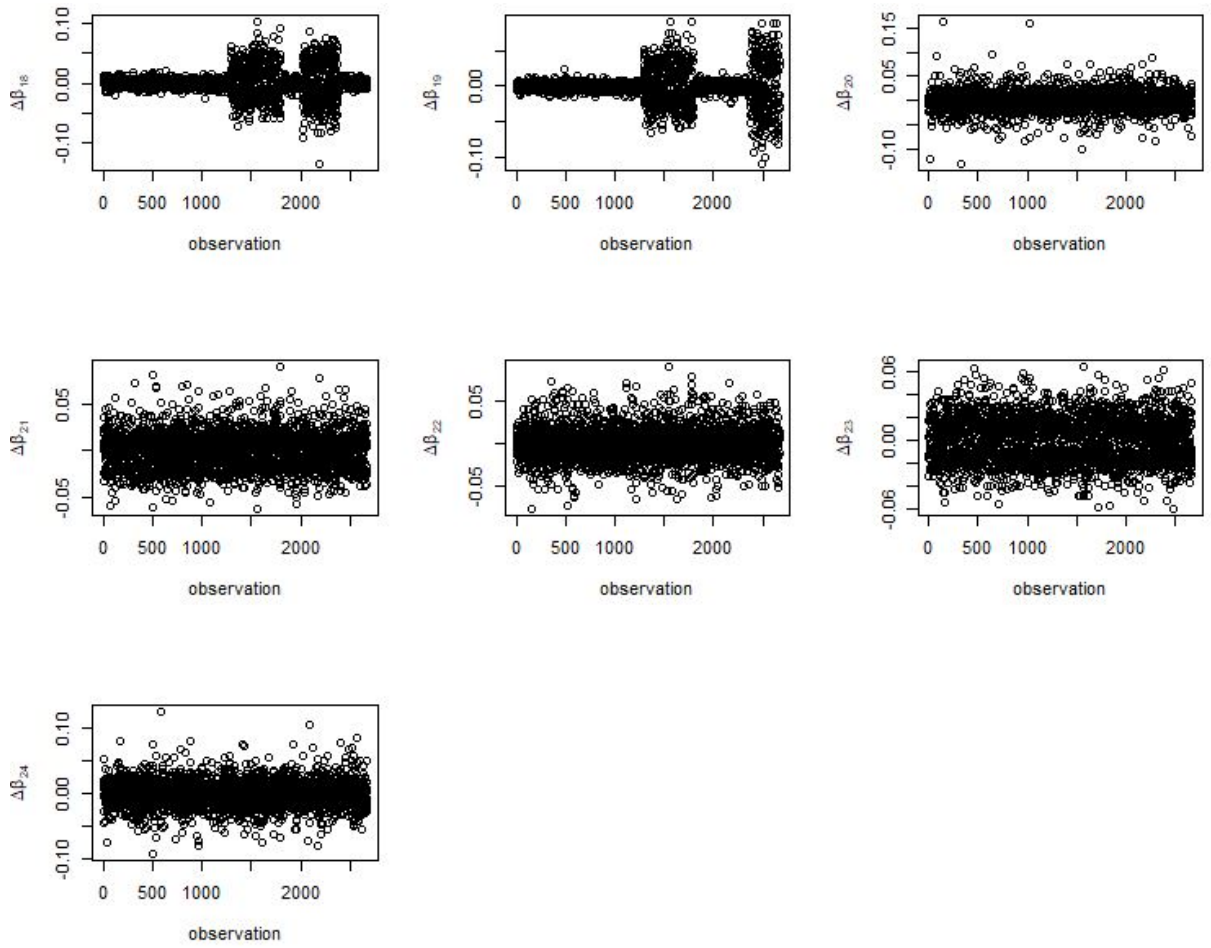

Fig. 9: Dfbetas for 3^{rd} wave, $\beta = 18, ..., 24$.