NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Modelling the Evolution of Domestic Violence Occurrences in Portuguese Municipalities

What Causes Domestic Violence?

Ana Clara do Carmo St. Aubyn

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

MODELING THE EVOLUTION OF DOMESTIC VIOLENCE OCCURRENCES IN PORTUGUESE MUNICIPALITIES

by

Ana Clara do Carmo St. Aubyn

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor / Co Advisor: Mauro Castelli / Maria Jordão

November 2021

ACKNOWLEDGMENTS

Throughout the writing of this study, I faced a lot of challenges that could not have been overcome without the great deal of support and assistance I received.

First of all, I would like to thank my supervisors, Professors Mauro Castelli and Maria Jordão for making the writing of this thesis possible and for all the support provided when things did not go the way I wanted them to.

Secondly, I would like to thank my father for his precious support and readiness to enlighten me when I was in doubt. You were a fundamental piece of this thesis.

I would also like to acknowledge my friends Sofia, Rita, David and Pedro, as well as my boyfriend for never letting me give up and for sharing the ups and downs of these two semesters.

Finally, but not least important, I would like to thank my mother, who understands me like no other and always calms me down when needed. Thank you for always listening to me during the process of writing this thesis even if you do not have a slight knowledge of econometrics.

ABSTRACT

Throughout the last years, domestic violence has been a widely discussed topic and an essential concern when building healthy communities. To fight the prevalence of this problem, its causes must be addressed. These causes can come from personal indicators, but they can also come from issues in society at large. It is important to shift the debate from micro-level to macro-level analyses, looking for structural factors in societies that contribute to the evolution of the number of domestic violence occurrences as these are the factors that can be addressed by regulatory bodies.

When modeling domestic violence one can envisage two types of possible explanatory variables: risk factors and protective factors. The first ones, as the term suggests, increase the risk of domestic violence, causing a high number of occurrences when very prevalent. Protective factors do the opposite, buffing the risk for domestic violence. The identification of risk factors is very important for the prevention of violence and to guide policies. However, identifying protective factors is of the utmost importance as their presence in Societies can be promoted to avert the occurrences.

The present document studies the evolution of domestic violence occurrences in the municipalities of the Portuguese mainland between 2009 and 2019 resorting to panel data analysis methods. The constant coefficients and the fixed effects approaches are employed to try and understand the relation between possible causes and domestic violence occurrences. While explaining a good proportion of the total variance encapsulated in the dependent variable was revealed to be a hard task, evidence of the importance of some variables in explaining domestic violence occurrences on a macro-level was found. These variables were the average number of children born to each woman in fertile age, the number of divorces *per* 100 marriages, the percentage of resident population with normal age for attending high school that is actually attending high school, the number of new marriages *per* 100 inhabitants, the percentage of men's monthly gain that women receive on average, the number of people enrolled in employment and vocational training centers *per* 100 inhabitants and the number of doctors *per* 100 inhabitants according to the Doctors' Professional Order. Most of these were shown to be risk factors, increasing the number of domestic violence occurrences.

KEYWORDS

Domestic Violence; Econometrics; Panel Data; Constant Coefficients; Fixed Effects; Risk Assessment; Explanatory Modelling

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AIC	Akaike Information Criterion
APAV	*Associação Portuguesa de Apoio à Vítima*
CC	Constant Coefficients
DGEEC	*Direção-Geral de Estatísticas da Educação e Ciência*
DGPJ	*Direção-Geral da Política de Justiça*
DVAM	Domestic Violence Against Minors
DVASA	Domestic Violence Against Spouse or Analogous
FE	Fixed Effects
GAM	Generalized Additive Model
GBV	Gender Based Violence
GER	Gross Enrolment Rate
IPV	Intimate Partner Violence
KNN	K Nearest Neighbors
OECD	Organization for Economic Co-operation and Development
OMA	*Observatório de Mulheres Assassinadas da UMAR*
RESET	Regression Specification Error Test
SC	Schwarz Criterion
SFI	Synthetic Fertility Index
WHO	World Health Organization
YDI	Youth Dependency Index

1. INTRODUCTION

Domestic violence is a widely discussed issue. According to Portuguese news agencies, the number of victims seems to be rising each year. Given this, it certainly is of the utmost importance to identify and address the causes of this problem. It is also true that the public, in general, is increasingly aware of the reality of domestic violence and this topic is becoming more relevant in the official media channels.

According to the yearly report[1] published by OMA – *Observatório de Mulheres Assassinadas da UMAR* – in 2017 there were 20 murders related to domestic violence in Portugal. Moreover, 28 cases of domestic violence were considered attempted murders. The report[2] from the following year states that the number of domestic violence related murders increased by 8, turning the reported number of murders related to domestic violence in 2018 into 28. The number of deaths related to domestic violence in 2019[3] was even higher than in the previous years, with 31 registered deaths. It is important to acknowledge that throughout the year of 2019 there were a total of 89 willful murders registered, as stated by the official statistics provided by *Direção-Geral da Política de Justiça* (DGPJ), making the previous number regarding murders related to domestic violence much more relevant and contextualized. The official report[4] about murder victims in 2019, released by APAV – *Associação Portuguesa de Apoio à Vítima* in June 2020, also stated that, regarding the 44 willful murders that they followed, 48% of those were perpetrated in a context of domestic violence. These numbers lift the veil on the sad reality Portugal is facing and clarify the need for addressing this problem.

1.1. THESIS OBJECTIVE AND RESEARCH QUESTIONS

The present dissertation aims to develop a model that allows an understanding of the causes of domestic violence in Portugal and explains, while quantifying the effect of each explanatory variable, the number of domestic violence occurrences. Considering the available data and its characteristics, the application of a panel data regression was selected as a viable solution for achieving the main goal.

During this research, the importance of several possible causes for domestic violence was tested. Since the objective was to explain the number of occurrences as well as possible, changes to some variables were considered during the process along with alternative models.

Keeping this in mind, the present dissertation proposes to answer the following questions:

- How did the number of domestic violence occurrences in Portugal evolve between 2009 and 2019?
- How well can panel data regression explain this evolution?
- What are the main causes of domestic violence?
- How does each explanatory variable affect the number of domestic violence occurrences?

[1] (Brasil, Alves, & Soares, Dados 2017, 2018)
[2] (Brasil, Alves, & Soares, Dados 2018, 2019)
[3] (Soares, Branco, & Alves, 2020)
[4] (APAV - Associação Portuguesa de Apoio à Vítima, 2020)

1.2. THE EVOLUTION OF DOMESTIC VIOLENCE IN PORTUGAL

To have a better understanding of the numbers, one can take a look at Figure 1.1. This figure represents the total number of domestic violence occurrences registered by police authorities in Portugal by year in the period between 2008 and 2019. From the figure, we can see a rise in the number of occurrences between 2008 and 2010, followed by a decrease from 2010 to 2012. In the period between 2012 and 2018, the number of occurrences remained relatively stable. However, one can witness a new rise from 2018 to 2019. The key point of this study is to find the causes for this evolution and explain their influence on the number of domestic violence occurrences.

Figure 1.1 - Domestic Violence Occurrences (on a national level)

In Portugal, under Article 152 of the Criminal Code, an aggression is categorized as domestic violence if the aggressor, repeatedly or not, inflicts physical or psychological ill-treatment, including physical punishment, deprivations of freedom and sexual offenses to:

- A spouse or ex-spouse.
- Someone from either the same or any other gender with whom the aggressor keeps or has kept a relationship analogous to that of spouses, even if without cohabitation.
- A parent of common offspring in the first degree.
- Someone who is particularly helpless, be it because of age, disability, illness, pregnancy, or economic dependency, with whom the aggressor cohabits.

Considering this, one can identify three categories of domestic violence as defined in the official statistics provided by DGPJ. The first one, which is called domestic violence against a spouse or analogous, includes all the topics mentioned above except for the last one. The second one, which is called domestic violence against minors, includes all aggressions to minors in which the aggressor cohabits with the victim. Finally, the last category, which is called others, includes the last of the topics mentioned above except for cases of domestic violence against minors. Figure 1.2 is a breakdown of Figure 1.1, splitting the total occurrences into these three categories. It becomes clear that most of the domestic violence occurrences in Portugal fit into the first category – domestic violence against a spouse or analogous – represented in the figure by the green line. The minimum value for this category was 20,394 in 2008, whilst the maximum value was 25,129 in 2010. The second most prevalent category is others, with a maximum value of 4,651 in 2011 and a minimum value of 3,083 in 2008.

Finally, the least represented category is domestic violence against minors, with a maximum value of 680 in 2008 and a minimum value of 430 in 2017.

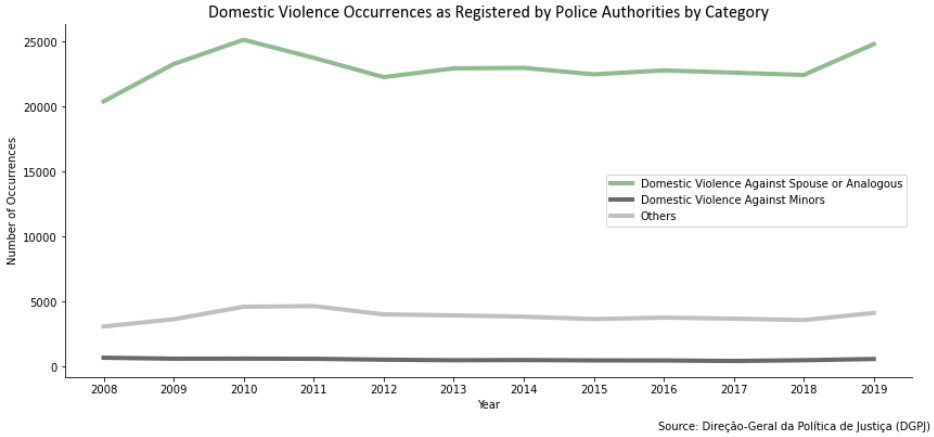


Figure 1.2 - Domestic Violence Occurrences by Category (on a national level)

Figure 1.3 shows the evolution of the first category only, allowing one to take a closer look at the evolution of the number of domestic violence against a spouse or analogous occurrences and notice that not only represents the majority of domestic violence occurrences in Portugal, but it also follows the pattern detected for the total occurrences described in Figure 1.1. This is the main category for domestic violence occurrences in Portugal and will also be the one used as a dependent variable (the variable to be explained) during the course of this study. This category will henceforward be referred to as DVASA.

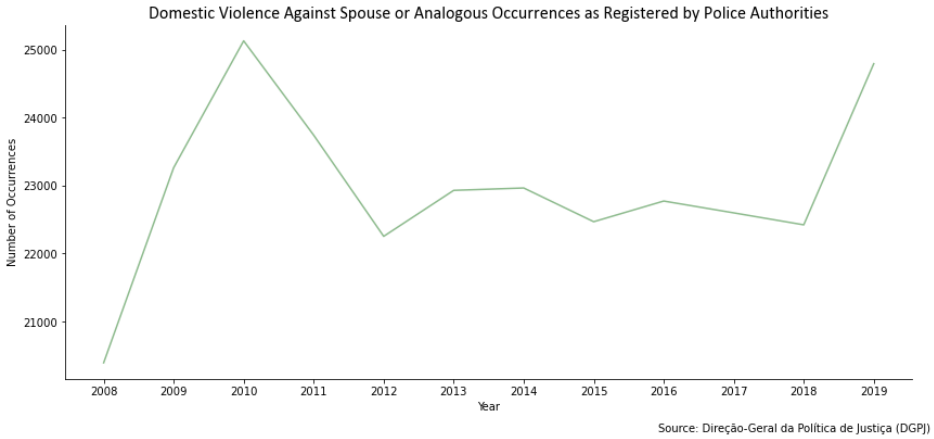


Figure 1.3 - Domestic Violence Against Spouse or Analogous Occurrences (on a national level)

2. LITERATURE REVIEW

Domestic Violence is not the only term used to describe this problem. Other popular terms include Intimate Partner Violence (IPV) and Intimate Partner Terrorism. Another common term in literature used to describe domestic violence perpetrated by men against women is Gender-Based Violence (GBV).

Domestic violence is a real issue that affects both countries as a whole and individual communities in multiple ways. A study[5] from 2018, measuring the global prevalence of IPV against women – which tend to be the most affected by this type of violence – combines data from 141 studies in 81 countries to show that, globally, 30% (this percentage was estimated using a 95% confidence interval) of women aged 15 or over have experienced some form of IPV. This percentage varies regionally. In Western Europe, where Portugal is located, it is estimated that around 20% of women aged 15 or over have experienced IPV.

In 2010, WHO (WHO - World Health Organization, 2010) expressed concerns regarding the lack of studies that provided information on whether observed factors are actual causes of violence. Most of the existing studies are based on cross-sectional population surveys and are from high-income countries, which leads to some uncertainty on whether the same factors associated with IPV in these countries are also associated with it in lower-income countries. Also, the existing studies are mostly focused on individuals rather than regions or countries, which leads to an identification of risk factors at individual or family levels, rather than at society or national levels. Unfortunately, this same problem was experienced when searching for literature for the present study.

(Amirthalingam, 2005) defines a spectrum of theories intended to explain domestic violence. At one end of this spectrum are theories that focus on the individual (related to psychological factors). Further on the spectrum are theories related to family and, finally, on the other end of the spectrum, are theories that focus on sociological factors that influence domestic violence. The first line of thought claims that the causes for violence can be internal (personality disorders, predisposition to violence, among others). Instead, family theories search for the cause of violence within the family unit, measuring most of the same variables as the individualists but this time in a family context. Finally, social structural theories of domestic violence shift the debate from micro-level to macro-level analyses, looking for structural factors in societies. The present study will follow this third line of thought, as this is the one that allows for public intervention and policy design, carrying the problem from the private arena to the public spotlight. Another theoretical model presented by WHO allows for the inclusion of all the factors mentioned above. It is called the ecological model and contemplates four levels of influence: individual, relationship, community and societal.[6] Since the present study focuses on the number of occurrences by Portuguese municipality it is not possible to include other factors besides societal ones.

To investigate the causes of domestic violence, one must analyze data regarding its evolution. Measuring the prevalence of domestic violence occurrences may be a hard task, as it is a sensitive topic. An article[7] by Ellsberg *et al.*, written in 2001, compares three studies on domestic violence in

[5] (Devries, *et al.*, 2013)
[6] Further explanation on this model can be found on (WHO - World Health Organization, 2010)
[7] (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001)

Nicaragua. Two of them are focused on urban areas of the country (León and Managua) and the remaining one is a national-wide Demographic and Health Survey that included other themes besides domestic violence. All of them are interview-based studies. When comparing the results of the studies, the authors of the article conclude that domestic violence occurrences tend to be underestimated when the source relies on self-reporting. This underestimation is not random, as it depends on numerous factors such as the number of individuals present in the room at the time of the interview or the way the questions are posed. In many other cases, this category of violence suffers from underreporting, as it usually happens in private spaces and the perpetrator is someone close to the victim. This makes it hard for the victim to come forward and for others to realize that something wrong is happening.

Information about domestic violence occurrences is almost always focused on women as the victims. Consequently, domestic violence against men tends to be concealed, as men are less likely to report such occurrences because of embarrassment, among other causes (Barber, 2008). This causes the number of domestic violence occurrences against men to be systematically underreported. The response by authorities is also at stake when it comes to men reporting domestic violence as the victim. (Barber, 2008) shows that, on a national-wide UK survey regarding domestic violence against men, only 3% of men who reported the occurrence to the police were taken seriously, whilst the remaining were threatened with arrest, ignored, or actually arrested. One can conclude that, due to the nature of the crime and to prejudices about the way society works, domestic violence occurrences are almost always underreported.

Healthcare workers are key players in detecting and listening to reports on domestic violence. (Cann, Withnell, Shakespeare, Doll, & Thomas, 2001) aims to find the healthcare workers (by gender, role and specialty) that have better responses and knowledge when it comes to domestic violence reporting and detection. This study includes healthcare workers from primary care, mental health, obstetrics and gynecology in the county of Oxfordshire, England. A questionnaire was presented to each of the workers and they were assessed according to a rating system on both knowledge and correct response. Women, nurses and mental health workers showed better results in dealing with domestic violence. It is possible that municipalities with more healthcare workers that fit into these categories show higher domestic violence rates, as occurrences can be reported by these workers. Many victims of abuse seek medical help, which makes it important for healthcare workers to have experience in handling these situations.

Domestic violence has been associated with many socioeconomic variables, such as the wealth and education of both the victim and the aggressor. It has been said that middle-level socioeconomic and well-educated groups tend to have the lowest prevalence of occurrences, whilst poorer groups tend to have the highest (Campbell, 2002).

Domestic violence can also be related to gender inequality. As explained in (Aizer, 2010), there are several theories regarding wage gender inequality and domestic violence against women. The one supported by this article states that, as the wage gap between genders decreases, women get more bargaining power and, consequently, domestic violence decreases as well. However, there are other possible hypotheses. The first one is the “male backlash” hypothesis which states that as the wage gap decreases, violence increases against women because aggressive men feel as if their traditional gender role might be threatened. The other hypothesis, the model of exposure reduction, claims that as the

wage gap decreases, the labor force participation of women increases and, consequently, domestic violence against them declines because women spend less time with violent partners. Either way, it is unquestionable that this is a variable of interest when it comes to justifying changes in domestic violence occurrences. One of the key points of this paper is that the relative or potential salary is more important in justifying violence patterns than the actual one. Taking this into account, it would be better to use a variable that reflects the potential wage of women vs. men instead of the actual wage gap. The results of this study, conducted in the state of California, United States, show that the decline in the wage gap between 1990 and 2003 accounts for nine percent of the decrease in domestic violence against women during that same period.

(Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001) mentions that some groups seem to be more at risk for domestic violence than others. It seems like women with more children are more prone to suffer assaults. Also, younger women were found to be at higher risk of violence. The more fragile the victims, the more likely they are to suffer from domestic violence.

Another factor that may be a possible cause of domestic violence is unemployment. A 2015 study[8] by Anderberg *et al*. focuses on the theory that a rise in female unemployment increases the number of domestic violence occurrences whilst a rise in male unemployment has the opposite effect. Using data from England and Wales, the authors show that this theory is well-grounded, revealing that a 1% increase in the male unemployment rate causes a decrease of 3% in domestic violence occurrences. A corresponding increase in the female unemployment rate has the opposite effect. Keeping this in mind, unemployment by gender may be a relevant variable to include in the present study.

When it comes to the Portuguese reality, APAV is the strongest association on the subject. According to an APAV report[9] published in 2018 that identifies the characteristics of victims of domestic violence in Portugal between 2013 and 2017, during this period this organization registered a total of 36.528 support processes in cases of domestic violence. In 31,317 of these processes, the victim was female (representing 85.73% of the total). Following the same line of thought, in 32,134 of the processes, the author of the crime was male (representing 85.93% of the total). This suggests that it might be useful to include some measure of the gender structure of the population as an explanatory variable for the present study. It is also mentioned in the same report that 41% of the victims were between 26 and 55 years old. However, another report[10] published by APAV also in 2018 that focuses on cases of domestic violence with male victims claims that the victim's age group with higher frequency is the 65 years or above one, representing approximately 28% of the processes included in the report. Figure 2.1 placed at the end of this chapter shows the percentage of processes for each age group (from 18 years to 65+) both in the general report (23,193 processes) and in the male-only report (2,745 processes). One can see that the patterns for each report are different, showing the importance of including the age structure of the population by gender in the present study.

Another key finding from the general APAV report is related to the family structure of the victims. Once again, it is shown that children are an important factor when determining the risk for domestic violence. From the 36,528 processes considered, 41.86% of the victims were in a nuclear family with children. Finally, it is also mentioned in the report that the marital status of the victim is also a relevant

[8] (Anderberg, Rainer, Wadsworth, & Wilson, 2015)
[9] (APAV - Associação Portuguesa de Apoio à Vítima, 2018)
[10] (APAV - Associação Portuguesa de Apoio à Vítima, 2018)

factor, as around 34% of the victims were married. This is a relevant percentage when compared to the 20.8% that were single, 16% whose marital status was unknown, 11.6% who were in a non-marital relationship, 8.7% who were divorced, 5.6% who were separated and 3.3% who were widowed. Therefore, including a measure of the number of married people may be relevant.

Still regarding the marital status, (Bowlus & Seitz, 2006) finds that women who are severely abused by their husbands are significantly more likely to divorce than women who do not face this problem. However, it is important to notice that women may be more likely to report violence in a past marriage than in a present one, causing an upward bias in this probability. This study uses data from all provinces of Canada, retrieved in 1993. The initial analysis of this data also revealed that women who experienced abuse by their partners tend to have lower levels of education and come from more violent backgrounds than women who did not face abuse. The same applies to the partners – husbands who abuse their wives tend to have lower levels of education. Another finding from this initial analysis is that women who are not working are more likely to face abuse, which meets the conclusions on (Anderberg, Rainer, Wadsworth, & Wilson, 2015). According to (WHO - World Health Organization, 2010), divorces may not only be a consequence, but also a cause of domestic violence. People who are separated or divorced tend to be more vulnerable, increasing their probability of becoming victims in a future relationship.

This same report by WHO lists some of the causes for domestic and sexual violence included in the literature they reviewed. Young age appears to be a risk factor for either becoming a victim of intimate violence or a perpetrator. It is foreseeable that populations with a higher proportion of young adults have higher rates of domestic violence occurrences. Lower levels of education are also consistently associated with both sides of the crime (victim and perpetrator). A higher level of education may act as a protective factor, since people with a higher level of education show lower levels of intimate partner violence. Another factor associated with both the victim and the perpetrator is poverty. Even though domestic violence cuts across all socioeconomic groups, people with lower incomes tend to be more at risk of becoming either a victim or an aggressor. One explanation for this, besides the hopelessness, stress and frustration caused by this condition, is the fact that shortage of money is a common cause for marital arguments and makes it harder for people to leave toxic relationships, as they are more financially dependent on each other. Some characteristics of the neighborhood may also influence the number of IPV occurrences, such as a lower proportion of women with higher levels of education, higher unemployment rates, a higher proportion of illiteracy and a lower proportion of women with high levels of autonomy.

(Ackerson, Kawachi, Barbeau, & Subramanian, 2008) examined the role of women's education and proximate educational context on GBV in India. A sample of 83,627 married women aged 15 to 49 years old from the 1998 to 1999 Indian National Family Health Survey was examined. The study considered that not only does the level of education of the woman herself influence her probability of becoming a GBV victim, but also does the general level of education of the community surrounding her. The results of this paper show that women with no education are 4.5 times more likely to report having suffered from domestic violence at some point in their life than women schooled for more than 12 years. Another relevant conclusion was that the probability for a woman who is living in the middle and lowest tertiles of female literacy to suffer from domestic violence at some point in her life was 1.18 and 1.10 times greater, respectively, than those of women living in the highest tertile neighborhoods. This study shows the impact that education has on GBV.

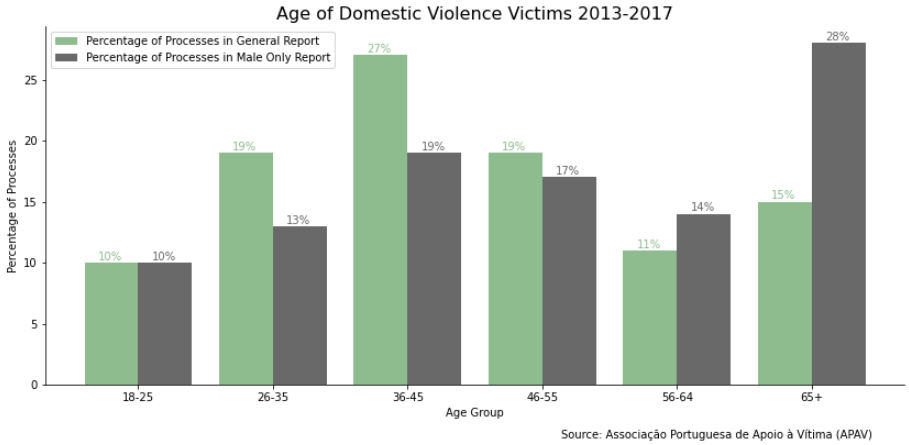


Figure 2.1 - Age of Domestic Violence Victims (2013-2017)

3. THEORETICAL BACKGROUND

The theoretical background of the present study was mostly written considering (Hill, Griffiths, & Lim, 2012) and (Wooldridge, 2013).

3.1. PANEL DATA

Data can be collected in multiple formats. The most widely discussed ones are cross-sectional data, pooled cross-sectional data, time series data and panel data.

Cross-sectional data is data collected for multiple units across the same period. Each observation represents a unit of the relevant population. This is the "common" dataset structure. When we combine cross-sectional data from different periods we create a pooled cross-sectional dataset. In this case, each observation represents a unit of the population in a specific period in time. It is not necessarily true that the same units are studied for the different periods. If we are studying the same unit across different periods in time, we create a time series. A time series shows the evolution of that unit through a specified time span. Finally, panel data, also called longitudinal data, is a combination of cross-section and time series data. Here, we have one time series for each included unit. The identifier of the unit and the period the data refers to are shown as variables in the dataset. The present study focuses on panel data, as there is one yearly discrete time series for each Portuguese municipality. Panel data analysis is a way of studying a subject in multiple sites periodically observed over a time frame.

Panel data may be considered short or long, balanced or unbalanced and fixed or rotating. A panel data is considered short when it studies many units for a short time period and it is considered long when it studies few units for a long time period. In the case of the present study, a short panel is being examined as it has 11 time periods and 278 units. A balanced panel has a number of observations equal to the number of units times the number of periods, meaning that all units are observed for all periods. If this is not the case, we have an unbalanced panel. The present study focuses on a balanced panel as all 278 municipalities are observed for all 11 years. Finally, if the same units are observed for each period, the panel data is fixed. If the set of units varies from one period to the next the panel data is rotating. In the case of this study, the 278 municipalities are fixed.

3.2. ECONOMETRIC PRIMER

Econometry starts with a theory about how some relevant variables are related to others. To express our ideas regarding these relationships we use functions. For most problems, it is not enough to know in which direction the variables are related (if they increase together, vary in opposite directions, etc.). Instead, one needs to know the intensity of that relation, which means how much a change in the value of one variable will affect the value of the other. To know these parameters of the relationship we create regressions.

Before generating an econometric model, one must keep in mind that relations among variables are not exact and, for this reason, no econometric model explains the exact behavior of its object of study. Instead, it describes the average or systematic behavior. This means that when presenting results from an econometric model one must always say "it is expected that…". Since the predictions are not exact, there is a difference between the actual value and the predicted value. This difference is the random

component of the model and is called the error term. It represents all factors that were not included in the model and includes the behavior uncertainty. The part that is explained by the model is the systematic component of the formula and it is decided by the investigator based on what the theory already existent states about the problem that is being studied.

One can say that the error term represents all things affecting the dependent variable other than the explanatory variables included in the model. It comprises the effects of relevant variables that were excluded from the model, measurement errors in both the dependent and explanatory variables, the effects of using a linear or any other form to generate results that do not follow that form and, finally, the natural randomness of observations.

To complete a model specification, the researcher must choose which variables to include and how to include them, keeping in mind the algebraic form of the relations. This form is also called the functional form and in the most basic scenario, it is assumed to be linear. An econometric model looks like the formula below where Y is the variable being studied (dependent variable), X_i are the variables that are assumed to affect the behavior of Y (explanatory variables), β_i are the coefficients of the explanatory variables that indicate how much they affect the dependent variable and, finally, e_i is the error term. β_0 is the expected value of Y if all explanatory variables are 0 and is called the constant of the model.

Systematic Component | Random Component

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$$

As it can be seen in the equation above that represents a multiple regression linear model, the model relates a dependent variable Y to a set of explanatory variables (X_1, X_2 and X_3) and to a random error term *e*. This remains true for other econometric models, changing the set of explanatory variables, the functional form, etc. The βs are the parameters that are estimated by the model and they can be interpreted to explain the relationships among variables. In the example above, one can say that a change in X_1 of one unit will represent a change of β_1 units in Y, holding everything else constant. Taking this into account, one can start to understand that the values calculated using an econometric model are, in fact, a conditional average. The predicted value of Y is the average of Y if the explanatory variables assume a set of fixed values. Let us consider the example of the present study: the probability density function for Domestic Violence Against Spouse or Analogous (DVASA) is illustrated in Figure 3.1 below, with a gray dashed line indicating the average value.

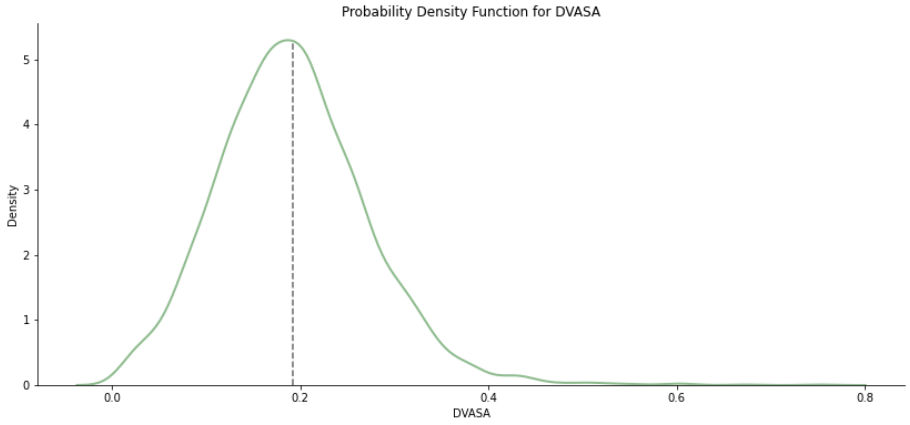


Figure 3.1 - Probability Density Function for DVASA

If we have a simple linear regression model, with only one explanatory variable (let us consider Education as the only explanatory variable), the predicted value for DVASA in a given point would be the average value of DVASA when Education attains a certain value. Therefore, we can calculate a conditioned probability density function and find its average to find the predicted value. Figure 3.2 below shows the plot for the probability density function of DVASA conditioned to when the variable GER (a measure of education – further explanation on Chapter 4.2) is at its most common value. One can see that the average value of this new distribution shifted slightly to the right when compared to the general distribution of DVASA. These are the differences that will be reflected in an econometric model.

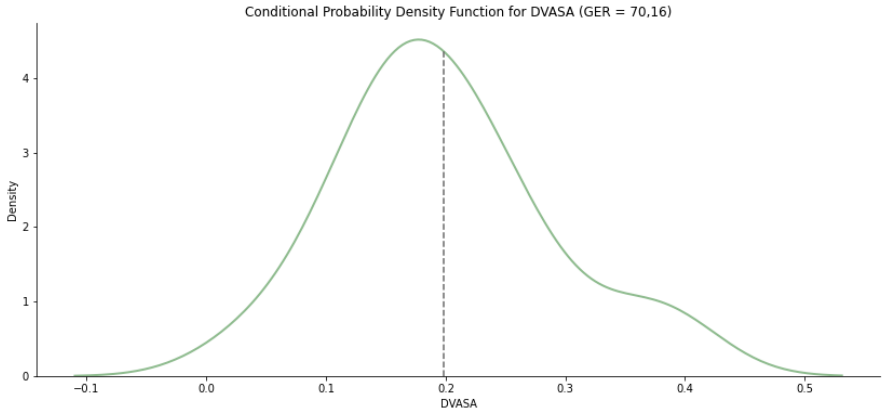


Figure 3.2 - Conditional Probability Density Function for DVASA (GER=70.16)

There are numerous types of econometric models. The simplest one takes only one explanatory variable and is called the simple linear regression model. If we include more than one explanatory variable, we are creating a multiple regression model. Econometrics also contemplates models for time series and panel data.

To understand how an econometric model is estimated the best option is to first study the simple linear regression model. In the previous example we considered DVASA as our dependent variable and GER as the only explanatory variable. We can plot the relationship between these variables to see how good of an approach we get. Figure 3.3 below shows a scatter plot of the two variables on the left. On the right side of the same figure, we added an approximation line to predict DVASA based on GER. However, adding a line is no guarantee that the expected value of the errors is 0, so the best approach is to use the Least Squares Estimators.

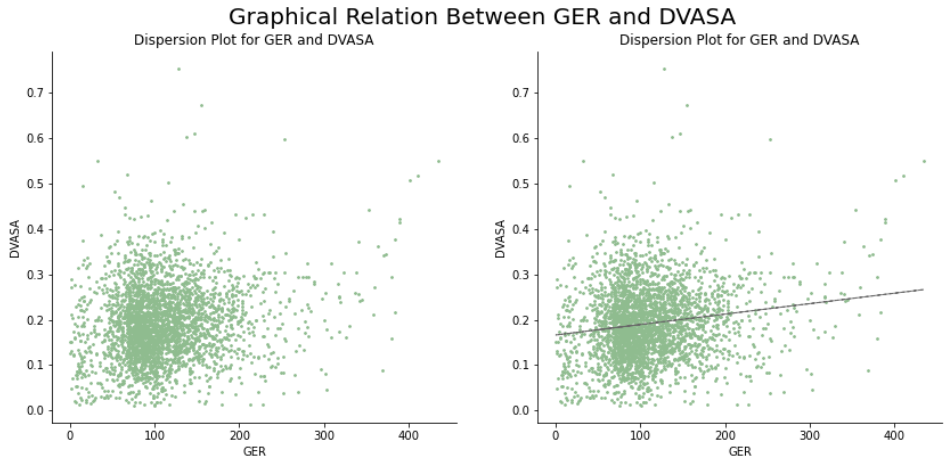


Figure 3.3 - Graphical Relation Between GER and DVASA

To estimate the intercept and the slope of the optimal line that describes the relationship between the dependent variable and the explanatory variable we want to make use of all observations available. The least squares principle says that we should add the line in a way so that the sum of squares of the vertical distances between the observations and the fitted line is minimized. It is important to square these distances so that positive distances do not cancel negative ones. These vertical distances are the residuals, hence the desire to minimize them. The sum of squares we wish to minimize is given by:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_{1i})^2$$

The function above is quadratic in terms of β_0 and β_1 and is shaped like a bowl. To find its minimum value we need to find the bottom of the bowl, which occurs where the slope of the bowl in the direction of each axis is 0. This is the same as saying that it occurs where the partial derivatives of S concerning β_0 and β_1 are 0. By calculating these partial derivatives, we obtain:

$$\frac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} = 0 \Leftrightarrow \beta_0 = E(y_i) - \beta_1 E(X_{1i})$$

$$\frac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} = \sum y_i X_{1i} - \beta_0 \sum X_{1i} - \beta_1 \sum X_{1i}^2$$

We can then replace β_0 as given by the first equation in the second one and obtain:

$$\sum y_i x_i - \left(E(y_i) - \beta_1 E(x_i)\right) \sum x_i - \beta_1 \sum x_i^2 = 0 \Leftrightarrow \beta_1 = \frac{E(y_i)E(x_i) - \frac{\sum y_i x_i}{n}}{E(x_i)^2 - \frac{\sum x_i^2}{n}}$$

These are the formulas for the least-squares estimators in the simple linear regression model and they can be used regardless of what the dependent or explanatory variables are.

Even though the simple linear regression model is the easiest to understand, most real-life econometric models take two or more explanatory variables, creating multiple regression models. Most of the conclusions regarding the simple linear regression model can be adapted for this type of model, except for some changes in the interpretation of the coefficients and in the degrees of freedom of the T distributions.

In a multiple regression model, we have several explanatory variables and we want to quantify the impact of each over the dependent variable, while controlling the effects of the remaining ones. This is the notion of *ceteris paribus* (see Chapter 3.3). Let us continue using the same example where DVASA is the dependent variable and GER is the explanatory variable, except that this time we are adding unemployment as another explanatory variable. We then obtain the following theoretical model:

$$DVASA_i = \beta_0 + \beta_1 GER_i + \beta_2 Unemployment_i + e_i$$

The βs in the above model measure the change in the dependent variable DVASA given a change of one unit in the respective explanatory variable, while holding everything else constant. The expected impact of an explanatory variable in the dependent variable is given by its partial derivate:

$$\frac{\partial y}{\partial x}$$

The objective in a multiple regression model is the same as it was in the simple regression model – to minimize the sum of squares of the vertical distances between the observations and the fitted line. The difference is that this time we need to work with matrices to find the least-squares estimators. In matrix notation, the model is given by:

$$Y = X\beta + e$$

In the previous equation Y is a matrix with only 1 column and n (n being the number of observations) rows containing the observed values for the dependent variable. X is a matrix with n rows and k+1 columns (where k is the number of explanatory variables). The first column in the X matrix only contains ones and will be used to calculate β_0. The remaining columns in X contain the observed values for the explanatory variables and will be used to calculate the respective βs. The matrix β has k+1 rows and only one column, containing the values of the least-squares estimators. Finally, the matrix e has n rows and only one column containing the residuals for each observation.

In this case, minimizing the sum of squares of the residuals is the same as minimizing the following (where the T means transposed):

$$\sum_{i=1}^{n} \hat{e}_i^2 = \hat{e}^T\hat{e} = \left(Y - X\hat{\beta}\right)^T\left(Y - X\hat{\beta}\right) = Y^TY - 2Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta}$$

To determine the values for the estimators that minimize the sum of squares we need to differenciate the above expression and equal it to 0, as we did with the simple regression model. We then obtain:

$$\frac{\partial(Y^TY - 2Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta})}{\partial\hat{\beta}} = 0 \Leftrightarrow \hat{\beta} = (X^TX)^{-1}X^TY$$

3.3. CAUSAL RELATIONSHIPS AND *CETERIS PARIBUS*

The goal in regression analysis is to find causal relationships between variables, that is, to determine whether a change in X causes a change in Y. To express our ideas regarding the relationships between variables we use functions. For example, to express a relationship between domestic violence and education one can write:

$$Domestic\_Violence = f(Education)$$

The previous function is a possible notation to say that the domestic violence occurrences in a certain place are a function of the education level of the people residing in that same place. This is the same as saying that the places where domestic violence occurs depend on the prevalent educational level in those places. However, the occurrences may not depend only on education, they may depend on many other factors as well. Keeping this in mind, to understand the relationship between domestic violence occurrences and education it is important to set aside the impacts of the remaining factors on domestic violence. The idea of *ceteris paribus* (c.p.) means to hold all other factors constant and is a key point in establishing causal relationships. Without holding the remaining variables constant one does not show that the change observed in *y* is caused by the change in *x*. This is also the reason why a simple correlation study is not enough to analyze causal relationships.

When we are studying the causal effect of *x* on *y*, the remaining variables that influence *y* are called the control variables. The reason to control for these variables is simple: we believe that *x* is correlated with other factors influencing *y*, which means that not holding these variables constant will make their

effects reflect on the coefficient for *x*. Since we feed data to the regression model, it is important to correctly determine the control variables that need to be held fixed. This is a critical part of regression analysis but may be hard, as usually not all factors influencing the dependent variable are observable or accessible. When one does not include an important control variable, its effects are reflected on the partial effects of other factors, making the latter incorrect.

Stating the difference between explanatory variables and control variables may be hard. Even if some variables can be considered control variables on all occasions, all explanatory variables are eventually control variables when it comes to explaining the partial effects of another variable. For panel data, the variables that determine the difference between observations (in the case of this study: municipality and year) are always control variables as, even if they can explain part of the variance in the data, their main goal is to distinguish between observations.

3.4. ECONOMETRIC ASSUMPTIONS

In every econometric study, there are at least two models: the theoretical model and the empirical one. The theoretical model describes a behavior but is an abstraction of reality. To convert the theoretical model into an empirical one, some assumptions must be made. These assumptions are very important because if they are verified our conclusions are warranted. However, if our model does not meet them, the conclusions may not be true.

The first assumption is that the variance of the dependent variable is a constant for each value of X. This means that in the case of the example using DVASA as the dependent variable and GER as the explanatory variable, for each value that GER may take, the average of DVASA may be different, but the measure of how much it varies around that average must be the same. This leads to a further assumption that the error term must have a constant variance as well. When this condition is satisfied, the data is said to be homoskedastic. When this condition is violated, the data is said to be heteroskedastic.

The second assumption is that the dependent variable is not only random but also statistically independent. This means that the value for one observation is not dep,endent on the value of another observation. In the DVASA example this means that the value of DVASA for one municipality is not dependent on the values for other municipalities on the dataset. Instead of assuming the independence of the variables, we often assume that the covariance of the variable is 0. Again, in the DVASA example, this means that if Y_j and Y_i are the values of DVASA for two different municipalities, $cov(Y_j, Y_i) = 0$.

The third assumption is also related to variance. Since the main goal of an econometric study is to understand how changes in the explanatory variables affect the dependent variable, it is important that the explanatory variables are scattered enough. Obviously, if there is no variance in the explanatory variables it is impossible to justify changes in the dependent variable from changes in the explanatory variables. Therefore, it is assumed that the explanatory variables take at least two values.

The fourth assumption is related to the error term of the regression. Since the total error of the model is the sum of the deviations between the actual value of an observation and the predicted value for it and the objective of a regression is to minimize these deviations, it is assumed that the expected value of the error ,term is 0. Furthermore, in a simple regression model it is also assumed that the expected

value of the error term given the explanatory variable is also 0. This is demonstrated in the equation below. It is important to remember that the expected value of the dependent variable given the explanatory variable is the predicted value so $E(y|x) = \beta_0 + \beta_1 X_1$.

$$E(e|x) = E(y|x) - \beta_0 - \beta_1 X_1 = 0$$

Finally, sometimes it is also assumed that the error term follows a normal distribution centered around 0. This is an optional assumption, but it is a strong one as the probability distribution of the parameters estimated in the regressions (βs) depends on the distribution of the error term which means that this is an important assumption for statistical analysis of the model. However, if there are enough observations one can base the conclusions on the central limit theorem that establishes that if the sample is large enough, the distribution of the sample means will be approximately normally distributed.

These assumptions exist so that we can determine the quality of the least-squares estimators. These estimators are supposed to be centered and efficient. An efficient estimator is one with minimal variance.

When the expected value of an estimator equals the real value of the parameter it is trying to estimate we say that we are dealing with a centered estimator. This means that the expected values of the estimators for the βs must be equal to the βs. In the simple regression model this would mean that:

$$E(\widehat{\beta_1}) = \beta_1 \quad and \quad E(\widehat{\beta_0}) = \beta_0$$

These equations are only true if the expected value of the error is 0, the error is not correlated with the variables and the data is coming from a random sample.

The variance of an estimator is also key to evaluating its reliability. It measures how much the values that the estimator may take are spread and, because of it, ends up measuring the precision of the estimator. Keeping this in mind, the smaller the variance of an estimator, the more precise it will be. For the simple regression model, when we calculate the variance of the estimators we get:

$$Var(\widehat{\beta_1}) = \frac{1}{\sum(x_i - \bar{x})^2}\sigma^2 \quad and \quad Var(\widehat{\beta_0}) = \sigma^2\left(\frac{\sum {x_i}^2}{n\sum(x_i - \bar{x})^2}\right)$$

Notice that σ2 stands for the variance of the error term and appears on both expressions. We can see that the larger the variance of the error term, the larger the variance of both estimators and, consequently, the more imprecise the estimation. We can also conclude that for the estimator to be efficient, the variance of the error term must be constant. Furthermore, we can see that the sum of the squared distances of x to its average is on both expressions on the denominator. This means that the larger the dispersion of the values in x, the smaller the variance of estimators, thus, the more precise they are. This means that choosing a sample that is diverse in terms of the values in the explanatory variables contributes to a higher precision of estimators.

An important theorem regarding the assumptions in an econometric model is the Gauss-Markov theorem. It states that if the fundamental assumptions are verified, the estimators for the βs are the ones with the least variance out of all the centered estimators. This means that they are the BLUE (Best Linear Unbiased Estimators).

When one is applying a multiple regression model, a new assumption must be verified. That is that none of the explanatory variables is a perfect linear combination of another. When this assumption is not verified, we are facing perfect multicollinearity and the least-squares estimators cannot be calculated.

3.5. STATISTICAL TESTS

When we estimate the coefficients for a regression, we are making a punctual estimate for the regression parameters. These estimates represent an inference over the regression model because after we calculate the values of the parameters for our sample, we intend to make an inference for the results to be applied to the population.

To make inferences we resort to interval estimation and hypothesis tests. Both these procedures are strongly based on the assumption that the residuals in the model follow a normal distribution. If this assumption is not verified, it is necessary to guarantee that the sample is large enough to assure that the least-squares estimators follow an approximately normal distribution through the central limit theorem. When resorting to this theorem the tests and intervals can be calculated, but their results are only approximate.

A hypothesis test allows us to evaluate the possibility that a parameter is equal to some value. In each hypothesis test, there must be a null hypothesis, an alternative hypothesis, a test statistic, a critical region, and a conclusion. The null hypothesis is denoted as H_0 and most of the time equals the parameter being tested to a specific value. It is the hypothesis that we intend to accept or reject according to the statistical evidence. Paired to any null hypothesis there is an alternative one, denoted as H_1. It is usually the contrary of the null hypothesis and is the one that is accepted if the null hypothesis is rejected.

To decide whether the null hypothesis is accepted or rejected, we must take into consideration the value of the test statistic. The distribution of this statistic is known if the null hypothesis is true but is unknown otherwise.

The critical region or rejection region of a hypothesis test depends on the form of the alternative hypothesis. It consists of the values that have a truly low probability of occurring if the null hypothesis is true. The logic behind this is that if the test statistic falls into the rejection region it is very unlikely that the null hypothesis is true. To define this region we must define a level of significance for the test. The level of significance of a test is the probability of rejecting the null hypothesis while it is actually true. This is also called a type I error. The level of significance is commonly designated by α and is usually either 0.01, 0.05 or 0.1. Another important concept when mentioning the significance of a test is the p-value. The p-value is the minimal level of significance with which we can reject the null hypothesis. If the p-value is less than the determined level of significance we reject the null hypothesis.

It is important to always remember while performing a hypothesis test that these tests are not able to prove that a null hypothesis is true or false. We can only conclude if the data is compatible with that hypothesis or not.

One of the most important hypothesis tests when it comes to econometric models is about the individual significance of the estimated parameters. For this, we test the hypothesis of a parameter being 0, which would mean that there is no significant relationship between the dependent variable

and the explanatory variable associated with that parameter. The hypothesis for this test (for a parameter β_k associated with an explanatory variable) are:

$$\begin{cases} H_0: \beta_k = 0 \\ H_1: \beta_k \neq 0 \end{cases}$$

In an individual significance test we use a t-statistic and a bilateral test. If the results lead us to reject the null hypothesis we can conclude that there is statistical evidence that the explanatory variable associated to the parameter being tested is a good determinant of the dependent variable. When presenting an econometric we should always indicate the p-value of the estimated parameters. For this, we commonly use asterisks that can be placed side-by-side with the estimate or below it. One asterisk means that the p-value was at most 0.1. Two asterisks mean that the p-value was between 0.01 and 0.01. Finally, three asterisks mean that the p-value was inferior to 0.01.

When we are estimating a multiple regression model it is relevant to perform joint hypothesis tests. In these tests, there are multiple restrictions in the null hypothesis. We use this type of hypothesis to test whether a group of variables should be included in the model, for example. To draw conclusions about these hypotheses we use a F-statistic. Let us imagine the following model:

$$DVASA_i = \beta_0 + \beta_1 GER_i + \beta_2 GER_i^2 + \beta_3 Unemployment_i$$

If we want to test the impact of the variable GER over DVASA we must perform a hypothesis test where the null hypothesis is $\beta_1=\beta_2=0$. To perform this test we must have a restricted model and an unrestricted model. Intuitively, the unrestricted model is the original model presented above, while the restricted model is one where the null hypothesis is true, so:

$$DVASA_i = \beta_0 + \beta_3 Unemployment_i$$

The joint hypothesis test in a F test that is based on the comparison of the residual sum of squares of both models. It verifies if the difference between residuals is large enough to justify the meaning of the parameters included in the test. If including the variables corresponding to the parameters being tested has a small effect on the residual sum of squares then we can conclude that, together, they are not an important contribution to the explanation of the dependent variable's variation.

If the null hypothesis for this test is true, the test statistic follows a F distribution with as many degrees of freedom in the numerator as the number of restrictions being imposed and n-k-1 degrees of freedom in the denominator (n is the number of observations and k the number of explanatory variables in the unrestricted model). As the F distribution is always positive the test is unilateral and the critical region is on the right side of the distribution.

There is a particular case of the joint hypothesis test in which we test all of the variables included in the model together. It is called the overall significance test. The logic is the same but the null hypothesis includes all the parameters and the restricted model becomes a model including only β_0.

As mentioned before, statistical inference relies on the residuals following a normal distribution. Even if the inference is valid with large samples due to the central limit theorem, it is desirable that this assumption is verified. For this, we can use the Jarque-Bera test. It is based on two measures, skewness and kurtosis. Usually, for a normal distribution, the value for kurtosis is 3 and for skewness is 0. The Jarque-Bera test compares the values obtained from the distribution of residuals to these values and

determines if the difference is big enough to reject the null hypothesis that the residuals follow a normal distribution.

3.6. HETEROSKEDASTICITY

When heteroskedasticity is present in a model it will negatively affect the statistical inference made around that model as the variance of the residuals is not constant and, thus, their distribution is not determined. Furthermore, it also means that the estimators for the coefficients are no longer efficient as their variance is not constant.

Heteroskedasticity can be detected through visual analysis, as seen in Chapter 3.8, Figure 3.5. However, there are multiple statistical tests for detecting heteroskedasticity. The first one is the Breusch-Pagan test, which is based on the function of variance. For this test, we create a model where the dependent variable is the squares of the residuals and the explanatory variables are the same as the ones in the original model we want to test for heteroskedasticity.

$$e_i^2 = \gamma_0 + \gamma_1 X_{1i} + \cdots + \gamma_k X_{ki} + \varepsilon_i$$

Then, we perform an overall significance test over this new model, which would mean that the null hypothesis would have all coefficients equal to 0. If we fail to reject the null hypothesis it would mean that there is no statistical evidence of the residuals being explained by the observed values for the explanatory variables. If the null hypothesis is rejected it means that the residuals are affected by the values in the explanatory variables and, thus, the model has a problem of heteroskedasticity.

One problem with the Breusch-Pagan test is that it assumes that the variables affecting the variance of the residuals are known. However, we might want to test the prevalence of heteroskedasticity without precise knowledge of the relevant variables affecting it. A possible solution is the White test.

The theory behind the White test is similar to the one behind the Breusch-Pagan test. We estimate a model where the dependent variable is the squares of the residuals and perform an overall significance test. However, this time we include not only the explanatory variables of the original model, but also their squares and, optionally, the interactions between them. So, for a multiple regression model with two explanatory variables this would be:

$$e_i^2 = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{1i}^2 + \gamma_4 X_{2i}^2 + \gamma_5 X_{1i} X_{2i} + \varepsilon_i$$

Once again, if we reject the null hypothesis of all coefficients being equal to 0, we diagnose the model with a heteroskedasticity problem. Nevertheless, there is still a problem with this test. Since we are including a lot of new variables in a model, we are losing a lot of degrees of freedom. We can see that for the previous example with a model containing only two explanatory variables, the White test model had 5 explanatory variables. To overcome this problem, we use a simplified version of the White test.

In the simplified White test, we use the prediction for the dependent variable obtained with the original model. We create a new model where the dependent variable is the squares of the residuals and the explanatory variables are those predictions and their squares.

$$e_i^2 = \gamma_0 + \gamma_1 \hat{y}_i + \gamma_2 \hat{y}_i^2 + \varepsilon_i$$

Once again, we perform an overall significance test and rejecting the null hypothesis means heteroskedasticity is present in the model.

3.7. Model and Variable Selection

One of the first steps in an econometric study is the choice of the model. This includes choosing which explanatory variables should be included and in which functional form should they be included. The wrong choice of variables may cause two different types of problems: the omitted variable problem or the irrelevant variable problem.

Omitting a relevant variable causes bias in the coefficient estimates for the remaining variables if these variables are correlated to the omitted variable. This happens because the effects of the omitted variables will be reflected in the coefficients of variables related to it. When the omitted variable has a positive effect on the dependent variable (a positive coefficient), the coefficients of variables positively correlated to it will suffer from an upper bias and the coefficients of variables negatively correlated to it will suffer from a downward bias. For cases in which the omitted variable has a negative effect on the dependent variable (a negative coefficient), the coefficients of variables positively correlated to it will suffer from a downward bias, while the coefficients of variables negatively correlated to it will suffer from an upper bias. If the omitted variable has a correlation coefficient of 0 with all remaining explanatory variables, their coefficients will remain unchanged and the effects of the omitted variable will be reflected on the error term of the model.

Including too many variables in order to avoid the omission problem is also not a good strategy, as the inclusion of an irrelevant variable reduces degrees of freedom in the estimation. Consequently, this causes larger standard errors for the coefficients and may lead to wrong conclusions. Another thing to keep in mind when it comes to including irrelevant variables is the possibility of spurious regressions. For example, the number of radishes harvested may explain some of the variability of the number of domestic violence occurrences, but this would be a completely random association.

One can conclude that it is important to choose an adequate set of explanatory variables to have a quality model. Unfortunately, this cannot be done by using a mechanical procedure that tells us the perfect set of variables. Instead, one should rely on the theoretical background already existent on the subject and on the result of statistical tests. Firstly, the investigator must choose the variables and respective functional forms using his understanding of the problem. Then a first model must be estimated and the coefficients and their magnitudes must be analyzed. If the coefficients have unexpected values, this may be caused by a misspecification. Then, hypotheses tests regarding the significance of variables on their own or as a set can be used, followed by some measures of selection such as the Akaike Information Criterion (AIC). Finally, the adequacy of a model can be tested using a general specification test called RESET. The procedure is always very subjective and is displayed in Figure 3.4 below.

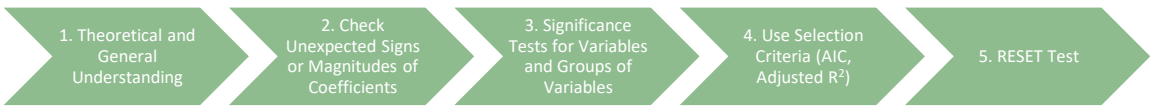


Figure 3.4 - Choosing Variables and Functional Forms for Models

The R^2 or coefficient of determination of a regression is a measure of the proportion of variability in the dependent variable that is justified by changes in the explanatory variables. Considering that the variability of the dependent variable has two components (one that is systematic and a random one), we can decompose it in the part that is explained by the model (explained sum of squares – ESS) and the part error that relies on the error term (residual sum of squares – RSS). The sum of these two results in the total sum of squares (TSS), which is equal to the variability of the dependent variable. Keeping this in mind, the formulas are as below. In these formulas $\bar{y}$ is the expected value of the dependent variable, $\hat{y}_i$ is the predicted value of the dependent variable for the i^{th} observation, n is the total number of observations and *e* is the error term.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = e^2$$

Since the objective of the R^2 is to inform about the proportion of variability in the dependent variable that is addressed by the model, it is calculated as $\frac{ESS}{TSS}$ or, equivalently, as $1 - \frac{RSS}{TSS}$. Since it is desirable that the model explains as much of the dependent variable's variability, it is desirable that the R^2 is as close to one as possible. However, the R^2 may not always be the best measure to compare models as it always becomes larger when adding new variables, even if the variables added are completely irrelevant. Keeping this in mind, the R^2 can only be used to compare models with the exact same number of explanatory variables. When the models have a different number of explanatory variables, the adjusted R^2 can be used as an alternative. This measure is calculated as following:

$$\bar{R}^2 = 1 - \frac{\frac{RSS}{(N-k)}}{\frac{TSS}{(N-1)}}$$

In the previous equation N represents the total number of observations and k represents the number of explanatory variables in the model. Using this measure, the value of the coefficient does not necessarily increase when a new variable is added because of the term N-k in the numerator. This means that when a new variable is added, the RSS decreases, but so does N-k, which causes the effect on $\bar{R}^2$ to be dependent on the amount by which the RSS falls. It is important to notice that the adjusted R^2 loses the interpretation of the R^2. However, we still want to maximize the adjusted R^2.

Another way of evaluating models is using information criteria. The two most used ones are the Akaike Information Criterion (AIC) and the Schwarz Criterion (SC). These criteria work in a similar way, having a first term that reflects declines in the RSS and a second term that work as a penalty for including too many variables. They work the opposite way as the R^2 or the adjusted R^2 in the sense that the objective is to minimize them. The criteria can be used when the number of explanatory variables in the model is different and their formulas are as following:

$$AIC = ln\left(\frac{RSS}{N}\right) + \frac{2K}{N} \qquad SC = ln\left(\frac{RSS}{N}\right) + \frac{K * \ln(N)}{N}$$

The Regression Specification Error Test (RESET) is designed to detect omitted variables and incorrect functional forms. Even when this test detects a misspecification error, it does not identify it. This test is applied after estimating the model and obtaining the predicted values for the dependent variable. These values are then used in a square and a cubic form as new explanatory variables for a new model. Then, a joint significance test is applied to test for the significance of the coefficients of $\hat{y}^2$ and $\hat{y}^3$. If

one concludes that these parameters are statistically significant it means that something is missing in the model and its specification can be improved.

3.8. MODELLING ISSUES

Sometimes the scale of the data is not the ideal. This can be altered without changing the underlying relationships between variables. For example, if the number of domestic violence occurrences by municipality was a really large number, one could decide to transform it by changing the unit of measure to hundreds of domestic violence occurrences. This would change the coefficients of the regression and their interpretation but would not change the results of statistical tests and other measures like the R^2. The process of changing the units of measure of a variable is called scaling the data.

Econometric models do not have to assume a linear relationship between the explanatory variables and the dependent variable. It is not always true that when an explanatory variable increases or decreases the dependent variable always continues to increase or decrease at the same constant rate. In order to include non-linear relationships in an econometric model one can transform the explanatory variables, the dependent variable or both. The most common transformations are powers of variables (quadratic or cubic, for example) and natural logarithms (ln). It is always important to notice that when applying transformations to any variables the interpretations of the coefficients affected by the transformation change as the slope of the line that defines the relationship is no longer constant and varies from point to point. The way the relationship is modelled is called the functional shape or form of the relationship. When choosing a functional form, the investigator must be aware of what theory tells about the relationship between variables and choose a shape that is sufficiently flexible to fit the data. Even if all considerations are taken, the functional form may still not be the best. One way of diagnosing this problem is by plotting the residuals against the values of an explanatory variable, for example. This is illustrated in Figure 3.5. If the points are randomly scattered around the plot and there is no evident pattern or trend it is probable that all model assumptions hold. This can be seen in Figure 3.5 on the left. If the residuals show increasing or decreasing variance, heteroskedasticity may be a problem (Figure 3.5 on the middle). If the functional form of the relationship is defined wrongly, the way the residuals are scattered around the plot may indicate the right one (Figure 3.5 on the right side).

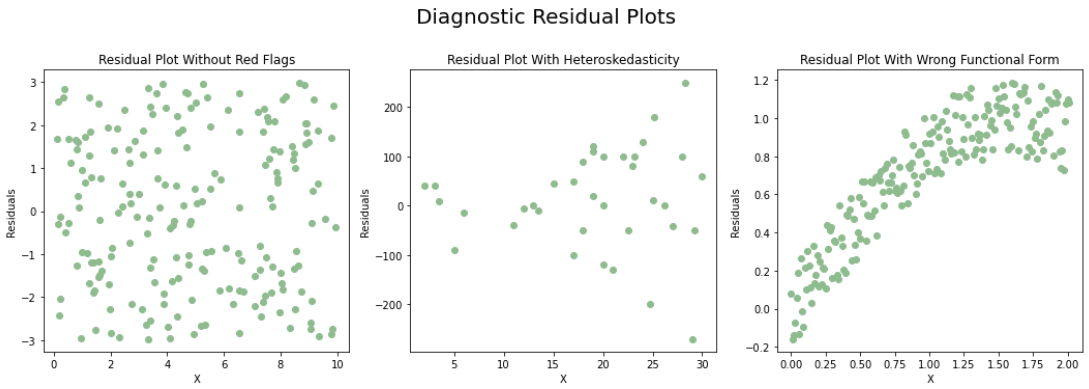


Figure 3.5 - Diagnostic Residual Plots (Example)

3.9. MODELS FOR PANEL DATA

The models estimated for panel data are obtained, generally, resorting to hypotheses about three components:

- The regression coefficients (β_{itr} where i is the index for the individual, t for the period and r for the variable);
- The explanatory variables and their relation to the error term;
- The stochastic properties of the errors.

One can distinguish between three types of models: the constant coefficients, the fixed effects and the random effects. The first one, the constant coefficients model, is used when there are neither significant cross-sectional effects nor significant temporal effects. In this case, one can undistinguishably use all the data and estimate an ordinary least squares regression. It is very unlikely that in a panel dataset there are no effects caused by the cross-sectional or temporal components, but they can be statistically not significant. If we are to consider three explanatory variables, a constant coefficient model can be written as following:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + u_{it}$$

In the previous equation one can notice the subscripts i and t. They denote the cross-sectional unit and the time period, respectively. In the constant coefficients model, as the name suggests, the coefficients β stay fixed for all units and periods (that is why they do not have subscripts). This type of model is also called a pooled model. If, besides assuming the common assumptions for the error term we also assume that it is uncorrelated over time and over the individuals, there is nothing that distinguishes this model from a multiple regression model.

The problem with the constant coefficients model is that it is unrealistic to assume that the errors for the same entity are not correlated. We then need to relax the assumption of zero correlation over time for the same entity. This does not mean that the error covariance is constant for the same entity, it just means that it can be nonzero. However, the covariance of errors for different individuals is still assumed to be 0. This leads to a problem of heteroskedasticity, which means that the least squares estimators are still consistent but their standard errors are wrong, leading to invalid statistical analysis. One way to overcome this problem is to use cluster-robust standard errors (the clusters are formed by the observations for each entity).

Fixed effects models allow the existence of individual characteristics of cross-sectional units by allowing different coefficients for different individuals. Keeping this in mind, a fixed effects model with three explanatory variables can be written as following:

$$y_{it} = \beta_{0i} + \beta_{1i} x_{1it} + \beta_{2i} x_{2it} + \beta_{3i} x_{3it} + u_{it}$$

In the previous equation one can notice the subscript i in the β coefficients of the model. This means that these coefficients can be different for each individual, creating what one can think of as a separate model for each cross-sectional unit. If using this line of thought, one can understand that this type of model can cause trouble on short panels, due to the degrees of freedom used when estimating it. If a panel has only five time periods and one tries to estimate a fixed effects model for each unit, one is only using five observations to create that model, which does not make sense and can even be

impossible if the number of coefficients to be estimated surpasses the number of observations for each unit. One way to avoid this problem is by using an alternative version of the fixed effects model, in which only the constant in the model differs from unit to unit. This can be written, considering three explanatory variables, as following:

$$y_{it} = \beta_{0i} + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + u_{it}$$

One can see that in the previous equation only the first β has the subscript i, meaning that all differences between individuals are assumed to be captured by the intercept in the model. One way to estimate this simple version of the fixed effects model is by crating a dummy variable for each unit that serves as an identifier for which individual the observation belongs to. By assuming the value 0 or 1 according to whether the observation refers to the unit or not, the corresponding coefficient will be ignored for observations that do not belong to the respective unit (by being multiplied by 0). In order to avoid problems related to multicollinearity, one unit must serve as the base for the others, which means that one unit does not have a corresponding dummy variable and is then represented by a constant in the model. The remaining coefficients represent the differences in the constant regarding the base unit for the remaining individuals. It is important to notice that, in short panels, the coefficients for the dummy variables are calculated using as many observations as the number of times periods contemplated for each unit, meaning that probably the central limit theorem does not apply anymore. Keeping this in mind, inferences on the coefficients for the dummy variables need normally distributed errors in order to be valid.

Fixed effects models can have different coefficients for the cross-sectional units, for the time periods or for both. The base idea is always the same – to reflect the fixed effects of each of these components differently.

4. DATA EXPLORATION

The dataset is composed by 19 explanatory variables and the dependent variable (DVASA occurrences). The methods used for calculating, treating, and standardizing all of these variables are explained in the next pages of this study. The dataset follows a panel data structure, having a column for the spatial dimension (Municipality) and one for the temporal dimension (Year). A summary of the existing variables can be found on Table 4.1, below:

Variable Name	Description
Municipality	Spatial Dimension of the Dataset.
Year	Temporal Dimension of the Dataset.
DVASA	Number of DVASA Occurrences Registered by Police Authorities by 100 Inhabitants.
Divorces	Number of Divorces for 100 Marriages in that Civil Year.
Elderly_Dependency	Number of People Aged 65 and Over for Every 100 People of Working Age, that is, Between 15 and 64 Years Old.
Female_Doctors	Percentage of Doctors Enrolled in the Doctor's Order Who Are Female.
Fertility	Average Number of Children Born for Each Woman in Fertile Age (Between 15 and 49 Years Old).
GER	Percentage of the Resident Population with Normal Age for Attending High School that is Actually Attending High School.
GER_Men	Percentage of the Male Resident Population with Normal Age for Attending High School that is Actually Attending High School.
GER_Women	Percentage of the Female Resident Population with Normal Age for Attending High School that is Actually Attending High School.
Marriages	Number of New Marriages per 100 Inhabitants.
Men65	Percentage of the Total Population of the Municipality that Represents Men With 65 Years or More.
Mental_Health	Percentage of Total_Doctors that are Specialized in Psychiatry according to the Doctor's Order.
Middle_Aged_Women	Percentage of the Total Population of the Municipality that Represents Women Between 25 and 54 Years Old.
Monthly_Gain	Average Gross Amount that the Employees in the Municipality Receive Every Month Including basic remuneration and Other Remuneration Paid by the Employer (Overtime, Holiday Pay or Premiums).
SS_Pensions	Number of Pensioners for Each Person who Cashes for Social Security. A

	Pension is an Amount Attributed Each Month to Someone in the Event of Disability, Old Age, Occupational Disease or Death.
Total_Doctors	Number of Doctors by 100 Inhabitants According to the Doctor's Order.
Unemployment_Female	Number of Women Enrolled in Employment and Vocational Training Centers per 100 Inhabitants.
Unemployment_Male	Number of Men Enrolled in Employment and Vocational Training Centers per 100 Inhabitants.
Unemployment_Total	Number of people Enrolled in Employment and Vocational Training Centers per 100 Inhabitants.
Youth_Dependency	Number of Children Under 15 Years Old for Every 100 People of Working Age, that is, Between 15 and 64 Years Old.
Wage_Gap	Percentage of Men's Monthly Gain that Women Receive on Average.

Table 4.1 - Dataset Variable Description

4.1. DEPENDENT VARIABLE

Three datasets containing information regarding the dependent variable were retrieved from the official statistics website by DGPJ on the 4th of March of 2021. One containing information about the number of domestic violence occurrences nationwide, one with this data split by districts and a last one with the data split by municipalities. All of them contained information regarding three categories (as explained in Chapter 1. Introduction): domestic violence against spouse or analogous (DVASA), domestic violence against minors (DVAM) and others. Finally, for all three datasets, the data was collected for the period between 2008 and 2019, due to data availability.

The Portuguese Criminal Code provides for and punishes the crime of domestic violence. Domestic violence assumes the nature of a public crime, which means that the criminal procedure is not dependent on a complaint by the victim, just a complaint or knowledge of the crime is enough for the Public Ministry to promote the process. Thus, in Portugal, the registered number of occurrences of domestic violence does not depend only on self-report by the victim. However, as it is a crime that commonly takes place in the privacy of a home, many cases may depend on self-report. The dataset obtained focuses on data registered by police authorities and, according to (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001), may suffer from underreporting, as it depends on self-report to some extent.

The data recorded for Portugal as a whole is a discrete time series. For each category there is a set of 12 observations recorded at uniformly spaced time values, in this case, years. This remains true for the data regarding districts and municipalities, except that, for the first case, there is one time series per category and per district, and for the second case there is one time series per category and per municipality.

The evolution of the number of domestic violence occurrences in Portugal can be seen above in Figures 1.1 and 1.2. It becomes clear by the analysis of these figures and of the descriptive statistics on Table

4.2 that domestic violence against spouse or analogous is the most prominent category out of the three. One can see that, between 2008 and 2019, the yearly average of domestic violence occurrences was 27,394. Considering the same period, the yearly average for the DVASA category was 22,977.8, a value that clearly shows how relevant this category is for the total domestic violence occurrences. The remaining categories have less significant yearly averages.

	DVASA	DVAM	Others	Total Occurrences
Std	1226.32	75.22	435.55	1612.36
Minimum	20394.00	430.00	3083.00	24157.00
Mean	22977.80	537.92	3879.17	27394.90
Maximum	25129.00	680.00	4651.00	30340.00
Q3	23382.80	599.00	4039.00	27877.00
Median	22851.50	515.50	3800.00	27155.00
Q1	22457.50	484.00	3647.75	26683.50

Table 4.2 - Descriptive Statistics for the Dependent Variable (on a national level)

When comparing the evolution of total occurrences in Portugal (Figure 1.1) with the evolution of occurrences for DVASA (Figure 1.3), one can detect the same patterns. By calculating the difference in the number of occurrences for subsequent years, it is possible to notice that DVASA occurrences almost always justify over half of the growth or decrease in the number of total occurrences. This is only not true for 2014, when the number of DVASA occurrences increased by 35 but the number of total domestic violence occurrences decreased by 48 due to a decrement in the other categories. Figure 4.1 shows exactly this.

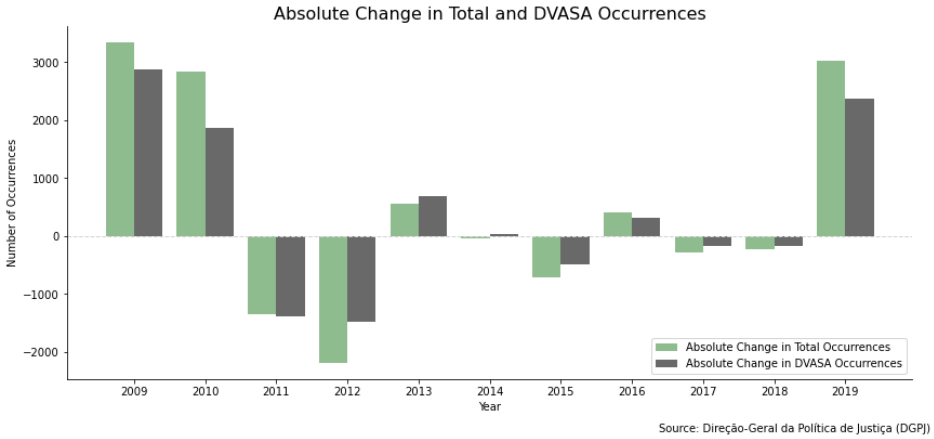


Figure 4.1 - Absolute Change in Total and DVASA Occurrences (on a national level)

Since the datasets only have 12 years worth of data, it would not be possible to perform a time series regression for Portugal as a whole, as there would not be enough degrees of freedom to provide powerful estimates. Keeping this in mind, a panel data regression will be performed with data regarding the years and municipalities. For this purpose, the dataset containing information about domestic violence occurrences by municipality must be analyzed.

Portugal is divided into 18 districts and 2 autonomous regions. Each of these is subdivided into municipalities. Currently, Portugal has 308 municipalities. The municipality data retrieved from the official statistics website by DGPJ measured the three domestic violence categories for the 308 Portuguese municipalities and for an extra N.E. one, meaning not specified (*não especificado* in Portuguese). Since it would not be possible to find the explanatory variables values for this special

case, this extra municipality was eliminated from the dataset. Furthermore, 12 of the 308 municipalities did not have values for all the categories. Corvo, the smallest island in the Autonomous Region of the Azores, only had data for the DVASA category. The remaining 11 municipalities (Pampilhosa da Serra, Golegã, Ribeira de Pena, Vila de Rei, Barrancos, Vila Viçosa, Penela, Alcoutim, Alfândega da Fé, São Roque do Pico e Aguiar da Beira) were missing data for the DVAM category.

The statistical confidentiality principle is stated in *Diário da República* (the Portuguese official gazette) in Law nº22/2008, the act that legislates on the National Statistical System. This principle, referred to in article 6 of the mentioned law, aims to safeguard citizens' privacy and secure trust in the Statistical System. Therefore, in the retrieved datasets, in order to respect the privacy of the people involved, numbers below 3 are not presented, being symbolized as missing values. Keeping this in mind, the number of missing values was calculated for each category. As one can see from Table 4.3, there were a total of 4,356 missing values among the three categories. The majority of these can be found in the DVAM and Others categories. The missing values for DVASA represent only around 2.3% of the total missing values in the dependent variable dataset.

Category	Number of Missing Values
DVASA	100
DVAM	2,862
Others	1,394
Total	4,356

Table 4.3 - Missing Values by Domestic Violence Category (Municipalities)

As mentioned before and seen on Figure 4.1, DVASA is the most prominent category in the total domestic violence occurrences in Portugal. Adding this to the facts that it is also the category with the least missing values (Table 4.3) and that it is the only category measured for all 308 municipalities, one can conclude that this is the best dependent variable for the present study.

In order to better understand the missing values and to find the best way to impute them, the difference between the national values for each year and the sum of the values for each municipality in each year was calculated (including the values for N.E.). This can be seen on Table 4.4. One can see that the number of missing values for the municipalities in each year is always very close to the number of occurrences reported on the national level but unreported on a municipal level. In order to address this, the missing values were replaced by the value 1 as it was considered better to keep information about municipalities with low occurrences than to remove them altogether from the study.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Missing Values	21	16	6	7	5	7	9	8	4	6	4	7
Difference	27	21	8	7	6	10	7	12	4	10	3	10

Table 4.4 - Missing Values for DVASA in Municipalities and Difference Between National Total and Municipality Total

It is important to keep in mind that the retrieved data was measured as an absolute value, which means that it did not consider the differences in the number of inhabitants for each municipality. Keeping the data as it was would have biased the future model, forcing it into thinking that the higher number of domestic violence occurrences in municipalities with the most population was caused by factors other than the number of inhabitants. In order to avoid this problem, the number of DVASA occurrences was

standardized according to the resident population in each municipality, as shown below. This way, the dependent variable is now the number of DVASA occurrences per 100 inhabitants.

$$DVASA_{standardized} = \frac{DVASA_{absolute}}{\left(\frac{Total\ Population}{100}\right)}$$

The data regarding resident population by municipality used to standardize the dependent variable was retrieved from the Pordata website on the 23rd of April of 2021. The definition of resident population in this case is the group of people who, regardless of being present or absent in a particular accommodation at the time of observation, lived in their usual place of residence for a continuous period of at least 12 months prior to the time of observation, or who arrived at their usual place of residence during the period corresponding to the 12 months preceding the moment of observation, with the intention of remaining there for a minimum period of one year.

4.2. EXPLANATORY VARIABLES

When modelling domestic violence one can contemplate two types of variables: risk factors and protective factors. The first ones, as the name suggests, increase the risk of domestic violence, causing a high number of occurrences when very present. Protective factors do the opposite, buffing the risk for domestic violence. The identification of risk factors is very important for the prevention of violence and to guide policies. Both types of factors can be divided into modifiable (for example education) and non-modifiable (for example gender and age) factors. The first ones are the most important when it comes to defining prevention policies, for logical reasons.

According to (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001) individuals belonging to families with more children are more prone to suffer assaults. As a way to include this factor in the present study, the synthetic fertility index (SFI) was considered as an explanatory variable. This index is the average number of children born for each woman in fertile age (between 15 and 49 years). In order for the generation renewal to be assured, the synthetic fertility index must be at 2.1. The data regarding this variable was retrieved from the Pordata website on the 15th of April of 2021 and included data from 2009 to 2019.

Another measure for the number of children is the youth dependency index. The data for this variable was retrieved from the Pordata website on the 6th of May of 2021. The youth dependency index is the number of children under 15 years old for every 100 people of working age, that is, between 15 and 64 years old. A value less than 100 means that there are fewer young people than people of working age. This variable had data for all the municipalities (without missing values) for the period between 2009 and 2019.

Healthcare workers play an important role in uncovering domestic violence occurrences and supporting the victims. According to (Cann, Withnell, Shakespeare, Doll, & Thomas, 2001), among healthcare workers, women, nurses and mental health workers tend to respond better to domestic violence cases. Keeping this in mind, data regarding the total number of doctors and the number of female doctors for each municipality was retrieved from the Pordata website on the 10th of May of 2021 in order to calculate the percentage of female doctors as below. This data covered the period between 2009 and 2019 and had missing values for some years in Pampilhosa da Serra, Oleiros and Lajes das Flores.

$$FemaleDoctors_{\%} = \frac{FemaleDoctors_{absolute} * 100}{TotalDoctors_{absolute}}$$

Following the same logic, the number of mental health workers (using psychiatry specialists as a proxy) was retrieved from the Pordata website on the 10th of May of 2021. The percentage of mental health doctors in the total of doctors was calculated as below. Once again, this variable had missing values for some years in Pampilhosa da Serra, Oleiros and Lajes das Flores.

$$MentalHealth_{\%} = \frac{MentalHealth_{absolute} * 100}{TotalDoctors_{absolute}}$$

Finally, the total number of doctors in each municipality was also used to create a variable that showed the number of doctors per 100 inhabitants of the municipality. This variable had data for the period between 2009 and 2019 and had no missing values.

As mentioned in the APAV report regarding male domestic violence victims (APAV - Associação Portuguesa de Apoio à Vítima, 2018), elderly men (65 years or more) tend to be more at risk. This can be seen on Figure 2.1. Taking this into account, the percentage of elderly men in the total population was included in the present study as an explanatory variable. The data regarding the absolute number of men with 65 or more years was retrieved from the Pordata website on the 23rd of April of 2021. This data was then converted to a percentage of the resident population as following:

$$Men65_{\%} = \frac{Men65_{absolute} * 100}{Total\ Population}$$

Still focusing on the elderly population, but this time without distinguishing between genders, the elderly dependency index was retrieved from the Pordata website on the 6th of May of 2021. The elderly dependency index is the number of people aged 65 and over for every 100 people of working age, that is, between 15 and 64 years old. A value less than 100 means that there are fewer elderly people than people of working age. This variable had data for all the municipalities (without missing values) for the period between 2009 and 2019.

Another way of measuring the level of dependency in a population is to consider the number of Social Security pensioners. A pension is an amount attributed each month to someone in the event of disability, old age, occupational disease or death. Data regarding the number of pensioners for each person who cashes for Social Security was retrieved from the Pordata website on the 11th of May of 2021. This data contemplated the period between 2009 and 2019 and had some missing values for the municipalities of Alenquer, Lagoa (Azores), Lajes das Flores, Santa Cruz das Flores and Corvo.

It is also mentioned in another APAV report regarding domestic violence victims in general (APAV - Associação Portuguesa de Apoio à Vítima, 2018) that most of the victims tend to be women with ages comprehended between 26 and 55 years. Keeping this in mind and with the same rationale as for the elderly men variable, the percentage of the population represented by women in these ages was included. The data regarding the absolute number of women between 25 and 54 years was retrieved from the Pordata website on the 27th of April of 2021. The boundaries of the age gap were as close as possible to the ones mentioned in the APAV report. However, they are not exactly the same as this data was not available. The percentage of middle-aged women was calculated as following:

$$MiddleAgedWomen_{\%} = \frac{MiddleAgedWomen_{absolute} * 100}{Total\ Population}$$

The same APAV report mentioned that a high percentage of the victims was married, showing that it might be relevant to include a measure of marriages as an explanatory variable for the present study. However, the number of marriages in a given year does not directly affect the number of domestic violence occurrences in that same year, as the marriage of the victims happens in years before the occurrence. Nevertheless, data regarding the number of marriages national-wide was retrieved from the Pordata website on the 28th of April of 2021 to test for correlations with the number of DVASA occurrences national-wide. When testing for the correlation between absolute values of DVASA occurrences and absolute values for the number of marriages the result was 0.48 which is neither a weak nor a high correlation. However, these values should be standardized according to the population. This standardization was made resulting in DVASA occurrences per 100 inhabitants and the same for the number of marriages. The correlation was then 0.31. Also, the evolution of both variables was plotted to check for common patterns (Figure 4.2) which were mostly not found. One can conclude that it might not be relevant to include this variable in the study. However, it might be interesting to check how this variable behaves when included in a regression and for that purpose data regarding the number of marriages by municipality was also retrieved from the Pordata website on the same date. This data was then standardized to reflect the number of new marriages per 100 inhabitants.

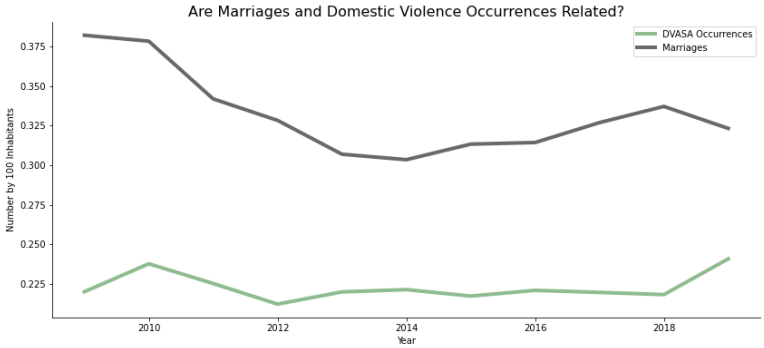


Figure 4.2 - Are Marriages and DVASA Occurrences Related?

Divorces are yet another controversial variable to include. However, they might be important as, if this variable works as expected according to (Bowlus & Seitz, 2006), it can be a good drive for action. That is, even though divorces are a consequence of domestic violence occurrences, they might be able to explain some of the expected underreporting in domestic violence occurrences. Also, if an increase in divorces is connected to an increase in domestic violence occurrences, the responsible entities can look for a rise in the number of divorces and, in that case, pay closer attention to domestic violence. (WHO - World Health Organization, 2010) mentions divorces as a cause for domestic violence and not only a consequence, as separated or divorced people tend to be more vulnerable, thus becoming more prone to being victims in a following relationship. Data regarding divorces was retrieved from the Pordata website on the 7th of May of 2021 in the form of divorces per 100 marriages. The data included the period from 2009 to 2019 and had some missing values, namely there were no values at all for Odivelas, no data for Castanheira de Pêra in 2013, no data for Barrancos in 2019, no data for Porto Moniz in 2016 and 2018, no data for Corvo in 2013, 2015, 2016, 2018 and 2019. This variable is calculated the following way:

$$Divorces\ per\ 100\ Marriages = \frac{Divorces\ in\ Civil\ Year}{Marriages\ in\ Civil\ Year} * 100$$

Another relevant variable according to (Campbell, 2002) is a measure of income, as the poorest strata of the population tend to witness more cases of domestic violence. Considering this, the monthly gain

of employees was included in the study. This refers to the amount that the employee receives every month. In addition to the basic remuneration, it includes other remuneration paid by the employer, such as overtime, holiday pay or premiums. It is calculated as a gross amount (before deducting any discounts). This data was retrieved from the Pordata website on the 26th of April of 2021, and it contemplates the period between 2009 and 2018. There was no information for this variable when it comes to all the 19 municipalities in the Autonomous Region of the Azores for the period between 2010 and 2013, which causes a total of 76 missing values.

Using the same dataset used for the monthly gain of employees, a measure of the wage gap between men and women was calculated. The data from Pordata, retrieved on the 26th of April of 2021, included the average monthly gain for all employees in a municipality as well as the average monthly gain for women only and for men. Once again, this data refers to the period between 2009 and 2018 and has no values for the municipalities in Azores for the period between 2010 and 2013. The variable here referred to as wage gap is the percentage of the men's monthly gain that women receive on average and was calculated as following:

$$WageGap = \frac{MonthlyGain_{women} * 100}{MonthlyGain_{men}}$$

According to (Anderberg, Rainer, Wadsworth, & Wilson, 2015) unemployment also influences domestic violence occurrences. Since the unemployment rate by gender was only available by regions and not municipalities, the number of people enrolled in employment and vocational training centers was used as a proxy. The values were calculated from a simple arithmetic average of the unemployed registered monthly in the employment and vocational training centers, so they are not always whole numbers. This data was retrieved from the Pordata website on the 29th of April of 2021 and had values for the period between 2009 and 2019. To test different possibilities, three variables were created from this data – female unemployment, male unemployment and total unemployment. All of them came in absolute values and had to be standardized by the number of inhabitants in the municipality. This standardization was done in the same way as the standardization of the dependent variable, resulting in the number of people enrolled in employment and vocational training centers by 100 inhabitants. It is also important to notice that there were no values regarding unemployment for the Autonomous Regions of the Azores (19 municipalities) and Madeira (11 municipalities), making it a total of 30 municipalities with no information.

Education may also play an important role in explaining the evolution of domestic violence occurrences. According to (Bowlus & Seitz, 2006), victims of violence tend to have lower levels of education and so do the perpetrators. Data regarding the gross enrolment rate (GER) was retrieved from the DGEEC – *Direção-Geral de Estatísticas da Educação e Ciência* – on the 23rd of June of 2021. The data was available for the period between 2003 and 2019 and did not have values for the municipalities in neither Azores nor Madeira. This indicator is calculated by DGEEC based on DGEEC enrollment data and INE (*Instituto Nacional de Estatística*) resident population data. It is calculated the following way:

$$GER = \frac{Students\ Enrolled\ in\ High\ School}{Resident\ Population\ With\ Normal\ Age\ for\ Attending\ High\ School} * 100$$

Regarding the GER, three variables were included: the total GER and GER by gender.

5. METHODOLOGY

The second step of this study, right after the data collection, was the data treatment and exploration, followed by modelling. Some of the data treatment was already described in the previous chapters, but the remaining part will be described in the present chapter. All calculations and plots were made using Python. The code used in the scope of this study can be found on a GitHub repository[11].

5.1. SERIES BREAKS

A lot of the explanatory variables had breaks caused by changes in the standards for defining and observing the indicator over time. According to the OECD (Organization for Economic Co-operation and Development) Glossary of Statistical Terms, "the specific causes of breaks in a statistical time series include changes in: classifications used, definitions of the variable, coverage, etc.".

The variables GER, GER_Women and GER_Men did not have series breaks.

Fertility, Youth_Dependency, Female_Doctors, Mental_Health, Men65, Elderly_Dependency, SS_Pensions, Middle_Aged_Women, Monthly_Gain, Wage_Gap, Unemployment_Total, Unemployment_Female, Unemployment_Male and Total_Doctors all had four breaks, all in 2013. Those were in Lisbon, Loures, Santarém and Golegã. The breaks in Santarém and Golegã are caused by the fact that the parish of Pombalinho was considered, from 2013 on, a parish belonging to the municipality of Golegã, no longer being a part of Santarém. Pombalinho is a small parish, with 7.7km^2 and 448 inhabitants, making this break neglectable. It was also a change in the parish configuration that caused the breaks for Lisbon and Loures. In 2013 a new parish called Parque das Nações was created, which included areas from both the municipalities of Loures and Lisbon. This parish is, from 2013 on, part of the municipality of Lisbon and it has 5.44km^2 and 21,025 inhabitants. Since only 34.2% of this area and 23.7% of this population belonged to Loures before the change, the effect of this break is neglectable.

The variable Marriages also had the four breaks for 2013 described in the last paragraph. However, it also had a break for each municipality in 2010 due to the fact that as of this year (inclusive), with the implementation of Law 9/2010 on the 31st of May, civil marriage between persons of the same gender became allowed. This last break cannot be considered neglectable and, so, data for the year of 2009 had to be removed from the variable in order to cancel the effects of the break.

Once again, the variable Divorces had the four breaks in 2013 related to the redistribution of parishes. However, it also had a break for all municipalities for the year of 2010 for the same reason the variable Marriages had a break in 2010 (Law 9/2010). From 2010 on (2010 included), divorces were allowed for persons of the same gender. Once again, this break cannot be considered neglectable and, so, values for the year of 2009 had to be removed for the sake of the coherence of the variable.

5.2. CORRELATIONS

It is important to know what the relation between variables within the dataset is. For this purpose, the correlation matrix was calculated using the Pearson Correlation Coefficient. The Pearson correlation is

[11] Text is underlined as it represents a link to https://github.com/anastaubyn/MasterThesis

used to evaluate the linear relationship between two variables. A relationship is linear when a change in one variable is associated to a proportional change in another variable. Spearman correlation evaluates the monotonic relation between rank variables or continuous ones and is usually used to evaluate the correlation between ordinal variables. In the case of the study, it makes more sense to use the Pearson Correlation Coefficient to evaluate associations among variables.

Since a panel data structure is being used, it can be considered a hybrid between time series and cross-sectional data. There is data for cross-section units (municipalities) and for periods (years). We can then unstack a single random variable into multiple random variables – considering the variable for the cross-section unit and the variable for the period. This allows us to calculate different types of correlations – one is a measure of association between cross-section units and the other is a measure of association between periods. The within cross-section correlations measure the association between the data in different periods for a given cross-section (in this case municipality) and is called contemporaneous correlation. To calculate it, one must group the data by period and calculate the correlation coefficients between cross-sectional units. The graphical views of the contemporaneous correlation matrixes for each year are presented in Annex I.

Figure 5.1. shows a graphical interpretation of the correlation among variables in the dataset. To build this correlation matrix, the average values for the correlations in each year were calculated, providing a summary of the information contained in the 11 yearly matrixes presented on Annex I.

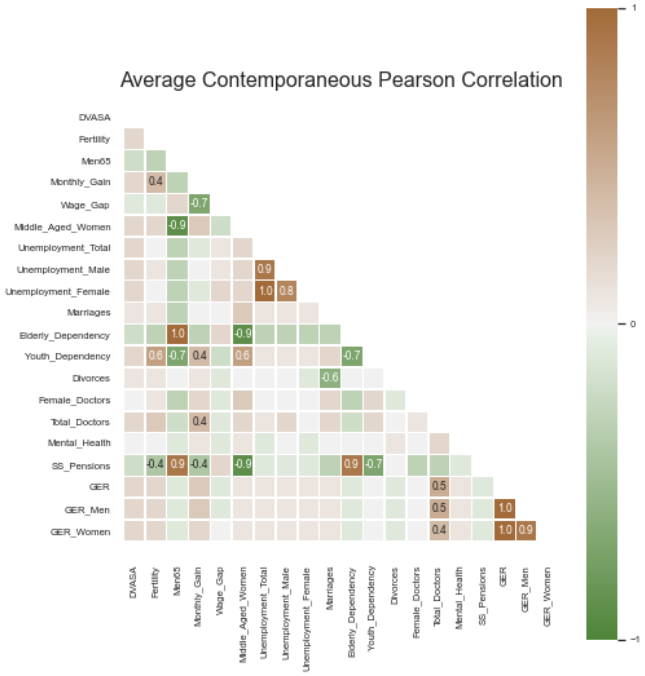


Figure 5.1 - Average Contemporaneous Pearson Correlation

Firstly, it is important to notice that both on the summary matrix plot and on the yearly matrixes plots only correlations with an absolute value above 0.4 are explicitly written. The first conclusion to take from the analysis of Figure 5.1. is the fact that there are no major correlations between any of the explanatory variables and the dependent variable. This can be due to the fact that the Pearson Correlation Coefficient measures linear relationships, meaning that it is possible that there still is association between DVASA and the remaining variables on a non-linear level. The second conclusion to take from the analysis of the heatmap is that there are some really high correlations between explanatory variables – Middle_Aged_Women and Men65, Unemployment_Male and

Unemployment_Total, Elderly_Dependency and Middle_Aged_Women, SS_Pensions and Men65, SS_Pensions and Middle_Aged_Women, SS_Pensions and Elderly_Dependency and, finally, GER_Men and GER_Women. All these correlations are almost perfect, with an absolute value of around 0.9. The are also some perfect average correlations among variables in the dataset – Unemployment_Female and Unemployment_Total, Elderly_Dependency and Men65, GER_Men and GER and, finally, GER_Women and GER. Most of these were expected, as they make perfect sense.

5.3. MISSING VALUES

Having many missing values lowers the quality of data. One can choose one of two approaches for dealing with missing values: either replace them or delete them. The method called listwise deletion consists of removing an entire observation from the dataset if any of its values are missing. However, this may weaken the statistical power of tests, as it reduces the number of observations and brings problems related to unbalanced panel data. Keeping this in mind, missing values were mostly replaced, as explained in the following paragraphs.

Some of the variables had missing values, but a lot of them were missing the entire year of 2008. Keeping this in mind, this year was removed from the study, making it a study about the evolution of DVASA occurrences between 2009 and 2019.

After the removal of the year 2008, the variables Fertility, Youth_Dependency, Middle_Aged_Women, Men65, Total_Doctors and Elderly_Dependency had no missing values.

The variable Female_Doctors had 22 missing values for the municipalities of Oleiros and Pampilhosa da Serra for the years between 2009 and 2014; Lajes das Flores for the years between 2009 and 2018. However, the variable Total_Doctors for all these observations was zero, meaning that these missing values were due to the calculation of the variable, when trying to make a division by zero. Since Female_Doctors represents the percentage of total doctors in the municipality who are female, if there are no doctors in the municipality this percentage is 0. These missing values were then all replaced by 0.

Mental_Health also had 22 missing values. Since this variable is also calculated by dividing by Total_Doctors, these missing values were in the same observations as the missing values for Female_Doctors. Once again, Mental_Health is calculated as a percentage of Total_Doctors so, if there are no doctors in the municipality, this percentage is also zero. These missing values were then all replaced by the value zero.

The variable Monthly_Gain had 384 missing values. Data for all the 19 municipalities in the Autonomous Region of the Azores for the period between 2010 and 2013 was missing. Besides that, there was no data at all for this variable for the year of 2019. The exact same observations had missing values for the variable Wage_Gap. Also, all three variables regarding education (GER, GER_Men and GER_Women) were missing all the data for both archipelago's municipalities. Keeping this in mind, the Portuguese archipelagos were removed from the study, making it a study about DVASA occurrences in the continental part of Portugal. When it comes to the missing values for 2019, since the trend will bias future models, it cannot be used to impute the values for this year. Taking this into account, plus the fact that there is now absolutely no data for both variables in 2019, data from 2018 was copied to 2019.

The variable SS_Pensions had 23 missing values. For 2009 in Alenquer, for 2010 in Lagoa (Azores) and then for the years between 2013 and 2019 for Santa Cruz das Flores, Corvo and Lajes das Flores. As Lagoa, Santa Cruz das Flores, Corvo and Lajes das Flores are all municipalities from the Autonomous Region of Azores, these observations were previously removed from the dataset. To impute the missing value for Alenquer in 2009 without using the trend, a KNN (K Nearest Neighbors) regression was applied. This method defines a neighbourhood of K observations that are the closest (using a chosen metric) to the observation being imputed according to the values of the other variables (or a subset of the other variables). Then, using that same neighborhood, it studies the relationships between the chosen subset of variables and the variable being imputed to create a regression. Finally, using the values from the observation being imputed, it fills in the missing value using the formula created from the neighbor observations. In this case, the subset of variables to use was chosen by calculating the Pearson correlation coefficient between SS_Pensions and the other explanatory variables for the observations for 2009. It was found that the correlation for this variable with Youth_Dependency was around -0.7, with Middle_Aged_Women was around -0.9, with Men65 was around 0.9 and, finally, with Elderly_Depency was around 0.9. These were then the 4 most correlated variables to SS_Pensions and were the ones used for the imputation. The metric used was the Euclidean distance. A neighborhood of seven observations was used and these were weighted according to the distance to the observation to impute.

The three variables related to unemployment (Unemployment_Total, Unemployment_Female and Unemployment_Male) all had 330 missing values, corresponding to the observations for all the 19 municipalities in Azores and all the 11 municipalities in Madeira for all the 11 years contemplated. Once again, this problem was solved by removing the Portuguese Autonomous Regions from the study.

Marriages had the break mentioned in Chapter 5.1 for all municipalities in the year of 2010, causing the removal of all values for 2009. Since using the trend will bias estimators in future models, it cannot be used to impute the values for this year. Keeping this in mind, plus the fact that there is now absolutely no data for this variable in 2009, data from 2010 was copied to 2009. The exact same procedure was applied to the variable Divorces for the exact same reasons.

After correcting the series break Marriages still had 10 missing values, all in Odivelas and for the period between 2009 and 2018.According to the metadata on this variable, the lack of data is due to the fact that the Civil Registry Office was not installed. The exact same thing happened in the variable Divorces. However, this variable had more missing values besides the ones for Odivelas. Data was missing for Castanheira de Pêra in 2013, Porto Moniz in 2016 and 2018 and Barrancos in 2019. The missing values for Marriages were imputed using the same technique as the one used for imputing SS_Pensions. The metric used was the Euclidean distance, observations were weighted according to this distance, the number of neighbors was seven and the variables used for the regression were determined according to the higher Pearson correlation coefficients – Elderly_Dependency, Middle_Aged_Women, Men_65 and SS_Pensions. Divorces was actually the highest correlated variable, but it would not make sense to use it as it is missing values for the same observations.

Finally, the missing values for Divorces were also imputed using a KNN regression. The metric used was the Euclidean distance, observations were weighted according to this distance, the number of neighbors was seven and the variables used for the regression were determined according to the higher Pearson correlation coefficients – Monthly_Gain, Fertility, Wage_Gap and SS_Pensions.

Lastly, the variables related to education also had some missing values. GER was missing data for 145 observations, Ger_Men for 151 observations and GER_Women for 154 observations. Again, the KNN regression technique was used. The metric used was the Euclidean distance, observations were weighted according to this distance, the number of neighbors was seven and the variables used for the regression were determined according to the higher Pearson correlation coefficients. All three variables regarding education are very similar (they are also very highly correlated), which ended up meaning that the variables used for imputing all of them were the same –Total_Doctors, Monthly_Gain and Fertility.

5.4. SUMMARY STATISTICS

As mentioned before the dataset is composed of data obtained for a period between 2009 and 2019 and for a total of 278 Portuguese municipalities (the Portuguese total of 308 except for the 19 municipalities in Azores and the 11 municipalities in Madeira). This makes for a total of 3,058 municipality-year combinations. Table 5.1. below shows the minimum, the median, the maximum, the mean, and the standard deviation (Std) for each of the variables contained in the dataset considering these 3,058 combinations. These are also called the overall statistics.

	Minimum	Median	Maximum	Mean	Std
DVASA	0.01	0.18	0.75	0.19	0.08
Divorces	0.00	64.30	700.00	70.46	40.11
Elderly_Dependency	14.20	37.45	97.9	39.31	13.03
Female_Doctors	0.00	50.00	100.00	47.15	15.87
Fertility	0.34	1.19	2.32	1.20	0.25
GER	0.50	100.00	434.90	110.55	54.78
GER_Men	0.90	94.60	429.20	106.59	54.95
GER_Women	1.80	105.80	623.10	115.97	58.05
Marriages	0.00	0.30	2.54	0.32	0.14
Men65	4.20	9.92	20.81	10.13	2.59
Mental_Health	0.00	0.00	33.33	0.88	2.31
Middle_Aged_Women	12.61	20.21	24.96	20.12	2.14
Monthly_Gain	616.60	867.50	2331.20	899.52	166.04
SS_Pensions	0.30	0.90	2.10	0.93	0.34
Total_Doctors	0.00	0.16	3.45	0.23	0.27
Unemployment_Female	0.61	2.45	7.68	2.58	1.01
Unemployment_Male	0.49	2.06	5.92	2.16	0.85
Unemployment_Total	1.18	4.55	12.23	4.74	1.79
Youth_Dependency	8.70	20.20	29.90	20.21	3.18
Wage_Gap	29.02	83.43	107.64	83.19	7.87

Table 5.1 – Overall Summary Statistics

Regarding the dependent variable, DVASA, we can conclude from the values on Table 5.1. that, considering all years and all municipalities, there are on average around 0.19 DVASA occurrences per 100 inhabitants. This value is very close to the median, which is 0.18. Both these indicators describe the center of the data, but extreme values (outliers) influence the mean more than the median. Close values for the median and the mean, when combined with not so high values for the standard deviation, lower the possibilities of outliers being present in the data. One can also see that the minimum value for this variable is 0 and the maximum is 0.75, meaning that the best municipality-year

combination (Pinhel – 2012) actually has 0 DVASA occurrences while the worst municipality-year (Mesão Frio – 2011) combination has 0.75 DVASA occurrences per 100 inhabitants. To further understand the distribution of the dependent variable, Figure 5.2. below shows the histogram for DVASA. From this figure, one can see that the distribution is skewed to the right, which might indicate the presence of upper outliers.

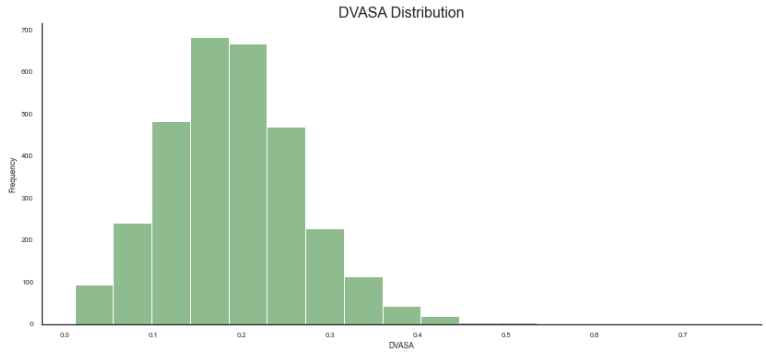


Figure 5.2 - Histogram for DVASA Distribution

Some of the values in Table 5.1. for the remaining variables indicate the presence of extreme values. For example, GER_Women has a median of 105.80 and a mean of 115.97 indicating that the data is probably skewed to the right. This can also be supported by the fact that the maximum value is very high (623.10) when compared to the central values and that the standard deviation (58.05) also suggests a large spread in the data. To evaluate the distributions of the explanatory variables visually, a grid of histograms is presented in Annex II.

Since the present study is dealing with panel data, two other types of statistics can be calculated to measure the dispersion of the data that is caused by differences between municipalities or within municipalities. When it comes to the first case, the between standard deviation is measured as the standard deviation of the municipality's average values. This measures the dispersion in the data that is caused by differences among municipalities. If the average values for the municipalities are very different from each other, the standard deviation will be higher, showing that there are significant differences among municipalities. In the second case, the within standard deviation is measured as the average of each municipality standard deviation. It measures the variation of data within each municipality, that is caused by differences caused by the years. The values for these alternative standard deviations are shown in Table 5.2. below.

	Overall	Between	Within
DVASA	0.08	0.06	0.05
Divorces	40.11	23.50	24.45
Elderly_Dependency	13.03	12.86	2.05
Female_Doctors	15.87	13.69	6.08
Fertility	0.25	0.19	0.15
GER	54.78	48.01	22.51
GER_Men	54.95	48.82	21.25
GER_Women	58.05	48.22	26.88
Marriages	0.14	0.11	0.07
Men65	2.59	2.55	0.48
Mental_Health	2.31	1.88	0.45
Middle_Aged_Women	2.14	2.09	0.43

Monthly_Gain	166.04	158.09	47.38
SS_Pensions	0.34	0.34	0.07
Total_Doctors	0.27	0.26	0.04
Unemployment_Female	1.01	0.83	0.56
Unemployment_Male	0.85	0.61	0.59
Unemployment_Total	1.79	1.39	1.12
Youth_Dependency	3.18	2.86	1.29
Wage_Gap	7.87	7.08	3.04

Table 5.2 - Overall, Between and Within Standard Deviations

Table 5.2. shows us that, generally, most of the variance is caused by differences between municipalities as it is true that, for almost all variables, the between standard deviation is larger than the within standard deviation. These results confirm an anticipated problem – the variables used do not have very dynamic evolutions from one year to another and are mostly static, which will make it difficult to justify differences in DVASA occurrences within municipalities.

5.5. Outlier Analysis

Outliers are extreme values that can be very prejudicial when building some models as they can distort statistical analyses and violate their assumptions. Extreme values increase variability within the dataset and decrease statistical power, meaning that removing them from the data can cause results to be more statistically significant. However, these values can bring a lot of information to the table and can even be some of the most relevant values in a dataset. Deciding whether to remove an outlier or not from a dataset can be mostly decided based on the causes for the presence of that value. There are three main causes for the existence of outliers in data: errors, sampling problems or natural variation.

In the first scenario, if an outlier is caused by an error, the most correct thing to do is to remove it or correct it. Errors can be spotted as impossible values for a variable, but they can also go unnoticed. In the present dataset, from the analysis of the histograms in Annex II there seem to be no outliers that are impossible values, but there is something suspicious. The distribution of GER_Women in much more skewed than the distributions of GER and GER_Men, suggesting a weird upper outlier. The value for GER_Women in Barrancos for 2010 is 623.1, which would mean that there were 6.231 times more female students enrolled in high school than those of expected age to attend it. This value by itself would not be impossible or suspicious at all, but when confronted with the value for GER_Men in the same municipality for the same year (4.3) both values become very odd. To further analyze these strange values, the evolution for all three GER variables in Barrancos is plotted on Figure 5.3 below. It becomes clear by the analysis of the plot that the value of GER_Women for 2010 is extreme and most likely an error. Thus, it was removed from the dataset and replaced with the value from the total GER in that same observation. No further suspicious values were detected.

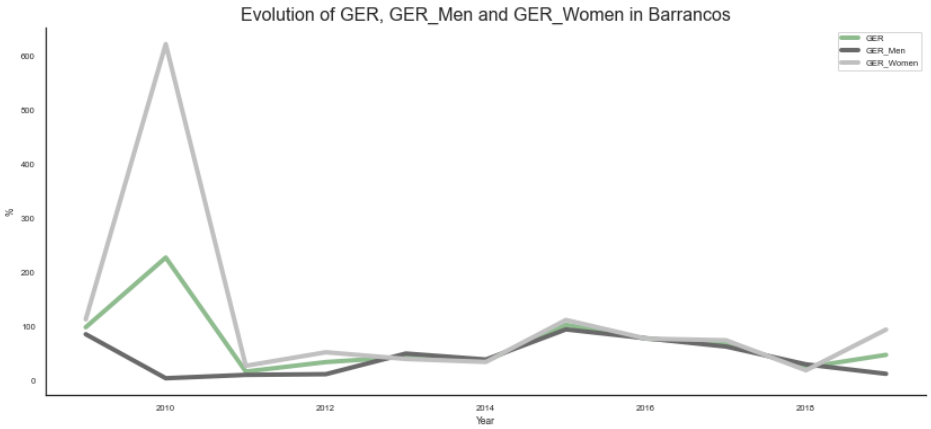


Figure 5.3 - Evolution of GER, GER_Men and GER_Women in Barrancos

The second scenario, sampling problems, is not applied to the present study, as it is studying a population rather than a sample. Outliers caused by sampling mistakes might occur when data is collected about an individual who does not belong to the target of the study. In this case, the outlier is also considered an error and should be removed from the dataset.

Finally, the last scenario is natural variation. These are the outliers that might be important for the study. Data distributions are centered around some point and spread from that point. This means that extreme values might occur but have a lower probability of happening. Even though these data points are unusual, they represent a natural part of the distribution and might add value to the dataset. For example, extreme values in one explanatory variable might justify extreme values in the dependent variable. For this reason, outliers apparently caused by natural variation were not removed from the dataset.

Besides the usual summary statistics, histograms and boxplots are common ways of exploring the presence of outliers. The histograms for both the dependent variable and the explanatory ones were presented in Figure 5.2. and Annex II before. However, for a better understanding of the distributions of all variables, a grid of boxplots is presented in Annex III. A boxplot is a summary of the data distribution where the sides of the central rectangle represent the first and third quartiles, the line in the middle of the rectangle represents the median and the “whiskers” represent a measure of 1.5 times the inter-quartile range. Theoretically, all data points that fall outside the “whiskers” of the plot are considered outliers.
One can see by the analysis of the boxplots that most of the supposed outliers fall very close to the distribution, as they are not far away from the end of the “whisker”. However, some variables have more extreme values: Divorces, GER_Women, Marriages, Mental_Health and Monthly_Gain. These extreme values may be important in justifying the variance in DVASA.

5.6. STATIONARITY

When dealing with time series data (panel data has multiple time series encapsulated inside of it) one can find time-dependent structures such as trend or seasonality. When these structures are present a time series is considered to be non-stationary, as the summary statistics do not remain fixed for all time periods. This causes variations in data that are caused by natural evolution of the numbers but that the model may try to capture anyway, adding bias to the results. Time series that are free of time-dependent structures are considered stationary. If the time series is stationary, the covariance

between two values of the series depends only on the amount of time separating those values and the summary statistics are fixed independently of the time period. This means that a stationary time series verifies the following conditions:

(a) $E(y_t) = \mu$;
(b) $Var(y_t) = \sigma^2$;
(c) $Cov(y_t, y_{t+s}) = \gamma_s$.

Since the present study is dealing with annual data, seasonality is off the table, as it is mostly found on monthly data, for example. However, trend may still be a problem. The two most common ways to detect the presence of time-dependent structures are visualizing plots of the variables of interest or using a Dickey-Fuller test. Changes in the variables over periods of time are important for visualizing whether a time series is stationary or not. If we have a variable y that is measured over some time periods t (y_t), the difference (Δy_t) given by $y_t - y_{t-1}$ is the change in the value of variable y from period t-1 to period t and is called the first difference.

The plots in Annex IV show the average (yearly mean values considering all municipalities) time series for all explanatory variables plus the dependent one side by side with the time series of the changes for the same variables. Since the mean and variance of a stationary time series are constant, its changes or first differences must fluctuate around a constant value, which does not seem to be the case for the plotted variables. However, as this is a result based on average evolution and changes it may not be the most reliable one. Keeping this in mind, it is necessary to resort to a more valid method, the Dickey-Fuller test.

The Dickey-Fuller test is based on the first order autoregressive model in which there are no external explanatory variables and the dependent variable is used as an explanatory variable for itself. This means that the dependent variable is related to past values of itself which means that each value of the variable contains part of the last period's value plus an error term. If the coefficient of the last period's value is less than one it means that the variable is stationary. This can be demonstrated as follows:

$$y_t = \beta_0 + \rho y_{t-1} + u_t \Leftrightarrow y_t - y_{t-1} = \beta_0 + (\rho - 1) y_{t-1} + u_t$$

One can see from the equation above that when ρ is less than one, the effects of y_{t-1} on y_t will decrease over time to the point in which a prediction for a period far from t will be unrelated to the value of y_t. However, if ρ is equal or more than one, the effects of y_{t-1} on the values of y for any period will never be annulated, which means that there are time-dependent structures and that the values of y rely heavily on the values of y for past periods. Keeping this in mind, the hypotheses for a Dickey-Fuller test are as following:

$$H_0: \rho \geq 1 \; vs. \; H_1: \rho < 1$$

It is important to note that the null hypothesis is that of the series not being stationary, meaning that if we fail to reject it, we are assuming non-stationarity. This causes a slight difference in the test statistic as non-stationary series have different properties altering the distribution of the usual t-statistic. To recognize this fact the statistic is called τ (tau) and its values are compared to specifically generated critical values.

A problem may arise when performing a simple Dickey-Fuller test as the error term may be autocorrelated. To avoid this, we should add as many lagged differences as needed. This number of

differences is determined by examining the autocorrelation function (ACF) of the residuals. This variant of the Dickey-Fuller test is called the Augmented Dickey-Fuller test and its equation is as following:

$$y_t - y_{t-1} = \beta_0 + (\rho - 1)y_{t-1} + \sum_{s=1}^{m} \beta_s (y_{t-s} - y_{t-s-1}) + u_t$$

The dataset that is the object of this study contains 5,560 time series (278 municipalities times 20 variables). The Augmented Dickey-Fuller test was applied to all of these series and it was possible to conclude that for 1,238 of them the null hypothesis was rejected with a significance level of 5%. If the significance level were to be pushed to 10%, the number of stationary series would be 1,502. Using either significance level it becomes clear that most of the series in the dataset are non-stationary. One way to avoid the problems caused by this condition is by removing the trend from the series. However, fixed effect models contemplate the possibility that the intercept may change over different time periods which can also be a solution. Finally, it is also possible to use differences as the variables. Nevertheless, since there is a much higher prevalence of the cross-sectional component in the present dataset than of the time series one this is not a severe problem. The present study is also working with short time series, making the power of Dickey-Fuller tests dubious, as it becomes hard to reject the null hypothesis even if it is false. One of the main problems of non-stationary series is that there is a danger of obtaining apparently significant results from unrelated data, which is called a spurious regression. However, the higher prevalence of the cross-sectional component avoids this problem.

5.7. MODEL ESTIMATION

Before estimating any models, it is important to understand how the explanatory variables and the dependent variable are related in terms of the functional form to be used. This can be analyzed visually with the aid of scatter plots. Keeping this in mind, a set of scatter plots showing the joint distribution of each explanatory variable with DVASA was plotted, as shown in Figure 5.4 below.

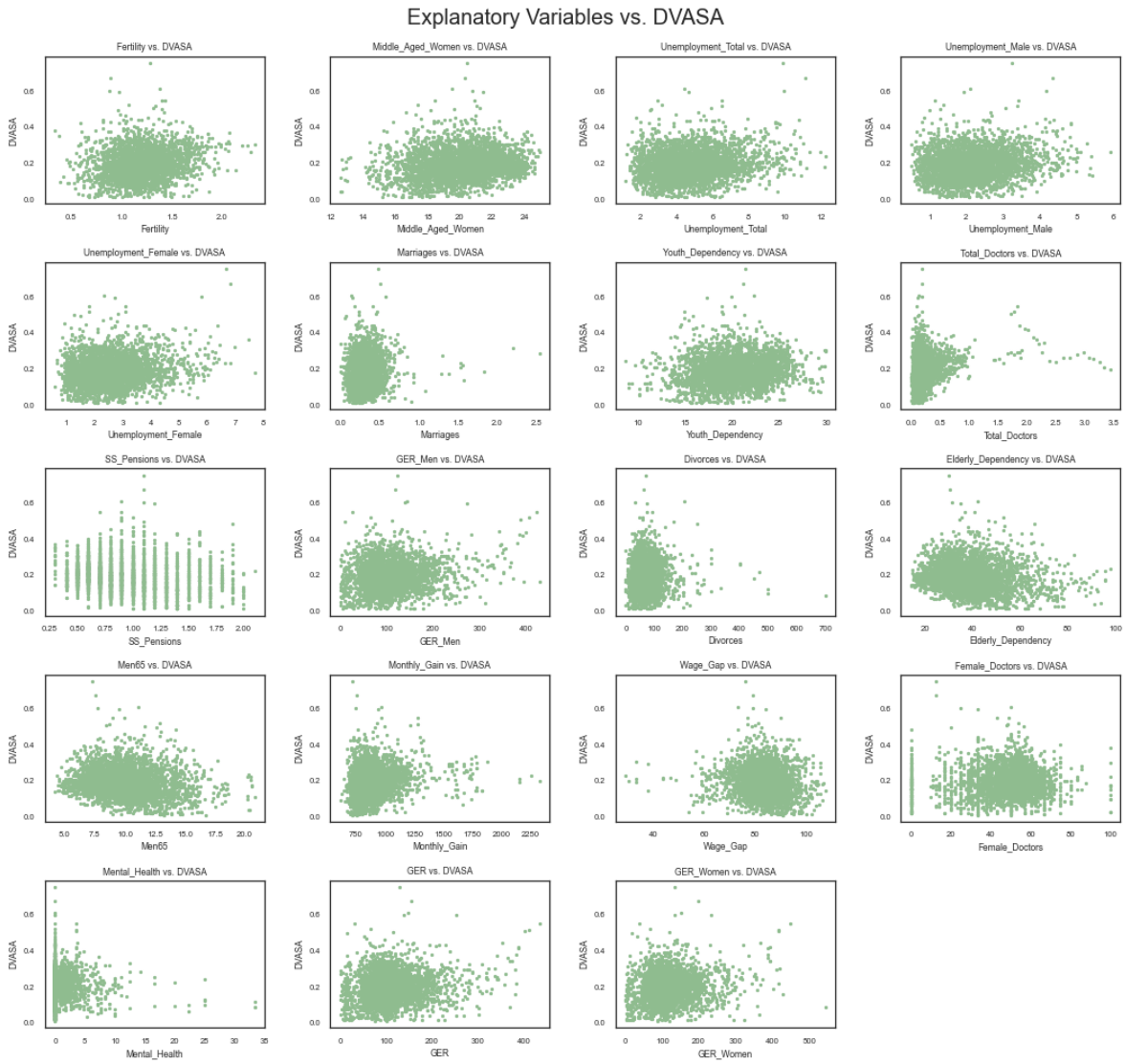


Figure 5.4 - Joint Distributions of Explanatory Variables with DVASA

From the analysis of the plots above, one can conclude that it might help to include some of the variables in different functional forms. Some explanatory variables, such as the ones related to unemployment and the ones related to education, exhibit a slowly increasing pattern, in a curvilinear form. This means that for higher values of the explanatory variable, the values of the dependent variable increase slowly. This type of relationship can be summed up through a log-log model (with the respective parameter being greater than one) or through a log-linear model (with the respective parameter being greater than 0). Either way, it might be beneficial to turn the dependent variable into a logarithmic form.

5.7.1. Model 1 (CC – Constant Coefficients)

To estimate the first model a constant coefficients approach was used. All 19 explanatory variables were used for this model in a simple linear form. From now on this model will be called Model 1. The theoretical model is as follows:

$$\begin{aligned} DVASA_{it} = {} & \beta_0 + \beta_1 Fertility_{it} + \beta_2 Men65_{it} + \beta_3 Monthly\_Gain_{it} + \beta_4 Wage\_Gap_{it} \\ & + \beta_5 Middle\_Aged\_Women_{it} + \beta_6 Unemployment\_Total_{it} + \beta_7 Unemployment\_Male_{it} \\ & + \beta_8 Unemployment\_Female_{it} + \beta_9 Marriages_{it} + \beta_{10} Elderly\_Dependency_{it} \\ & + \beta_{11} Youth\_Dependency_{it} + \beta_{12} Female\_Doctors_{it} + \beta_{13} Total\_Doctors_{it} \\ & + \beta_{14} Mental\_Health_{it} + \beta_{15} SS\_Pensions_{it} + \beta_{16} GER_{it} + \beta_{17} GER\_Men_{it} \\ & + \beta_{18} GER\_Women_{it} + \beta_{19} Divorces_{it} + u_{it} \end{aligned}$$

Model 1 was estimated using cluster-robust standard errors that allow for a relaxation of the assumption regarding the covariance of the error term. This means that the error for observations of the same municipality may be correlated but the covariance of error terms for different municipalities must be 0. The R-squared for this model was 0.1342 meaning that 13.42% of the variability in DVASA is represented by the set of explanatory variables chosen for this model. The result of the overall significance test for Model 1 was very positive, with a p-value of 0. Table 5.3 sums up the parameter estimates and their significance for this model.

	Parameter	Standard Error	T-stat	P-value
Constant	0.4115	0.1620	2.5409	0.0111
Fertility	0.0282	0.0105	2.6893	0.0072
Men65	-0.0103	0.0065	-1.5837	0.1134
Monthly_Gain	0.0000413	0.000027	1.5278	0.1267
Wage_Gap	-0.0003	0.0004	-0.6158	0.5381
Middle_Aged_Women	-0.0088	0.0044	-1.9877	0.0469
Unemployment_Total	4.5520	2.2377	2.0342	0.0420
Unemployment_Male	-4.5526	2.2369	-2.0352	0.0419
Unemployment_Female	-4.5356	2.2386	-2.0261	0.0428
Marriages	0.0368	0.0186	1.9731	0.0486
Elderly_Dependency	0.0010	0.0012	0.8918	0.3726
Youth_Dependency	-0.0031	0.0013	-2.3648	0.0181
Female_Doctors	-0.0000685	0.0002	-0.3551	0.7226
Total_Doctors	0.0393	0.0178	2.2053	0.0275
Mental_Health	-0.0006	0.0011	-0.6031	0.5465
SS_Pensions	-0.0460	0.0225	-2.0474	0.0407
GER	-0.0008	0.0005	-1.7929	0.0731
GER_Men	0.0006	0.0003	2.0282	0.0426
GER_Women	0.0004	0.0002	1.5352	0.1248
Divorces	0.0002	0.0000647	2.6329	0.0085

Table 5.3 - Parameter Estimates for Model 1

The last two columns of Table 5.3 show the results of the individual significance test for each parameter. P-values that exceed the significance level of 5% are highlighted in red. In this test, if the null hypothesis is rejected, one can say that there is statistical evidence that the parameter is not 0,

which means that the corresponding variable has an actual impact on the dependent variable. Considering a level of significance of 5%, one can see from Table 5.3 that there is no statistical evidence of some of the parameters being relevant to the model. It is the case of the parameters corresponding to GER_Women, GER, Mental_Health, Female_Doctors, Elderly_Dependency, Wage_Gap, Monthly_Gain and Men65. If one were to consider a significance level of 10%, only GER would be removed from the previous list. In future models this conclusion will be considered.

In order to test for heteroskedasticity in Model 1, a graphical approach was used first, plotting the residuals against the predicted values. If any pattern is detected in the plot (increasing or decreasing variance), there is heteroskedasticity in the model. The plot is shown in Figure 5.5.

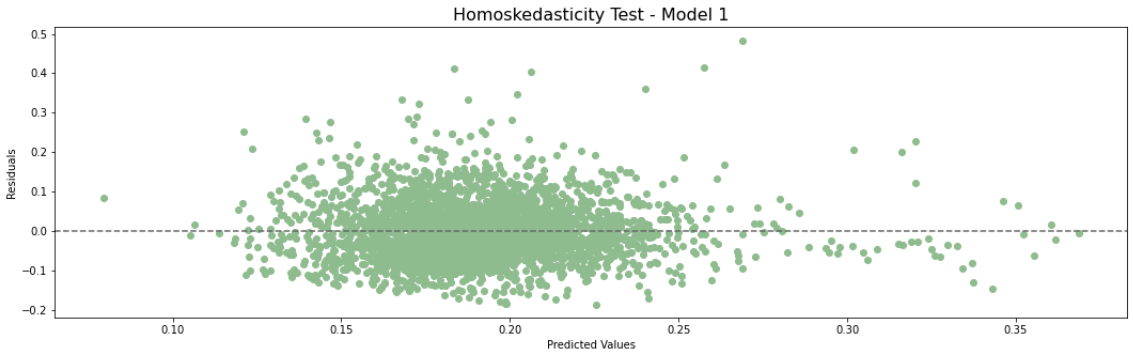


Figure 5.5 - Heteroskedasticity Test for Model 1

According to Figure 5.5, there does not seem to be any evidence of heteroskedasticity in the model. However, to guarantee this result, a formal test must be made. The Breusch-Pagan test is meant to detect heteroskedasticity. It takes the squares of the residuals created by the model and makes a regression on it with the same dependent variables as the original model. Then, an overall significance test is performed to see if all the coefficients can be 0. The results for this test show a really low p-value (1.27825047691774e-16), indicating that there is statistical evidence that at least one of the coefficients is not equal to 0, which means that there is heteroskedasticity in the model. Consequently, the estimators are not efficient. However, since cluster-robust standard errors were used, the results of statistical tests can be evaluated. No other tests on heteroskedasticity were performed for Model 1 due to the extremely low p-value for the Breusch-Pagan test.

In order to test for autocorrelation among residuals, the Durbin-Watson test was applied. The Durbin-Watson statistic is calculated using the residual sum of squares and the sum of squares of differences between consecutive residuals. It takes a value between 0 and 4 where the middle value (2) indicates no autocorrelation. Values between 0 and 2 indicate positive autocorrelation and values between 2 and 4 indicate negative autocorrelation. The statistic for Model 1 was 1.8941, which indicates a small level of positive autocorrelation.

Considering the results of the two tests, it might be a better option to use a fixed-effects or a random-effects model. Nevertheless, the plots of the residuals versus the explanatory variables for which the individual significance test showed statistical evidence of importance were plotted to check for wrong functional forms. These plots can be seen on Figure 5.6.

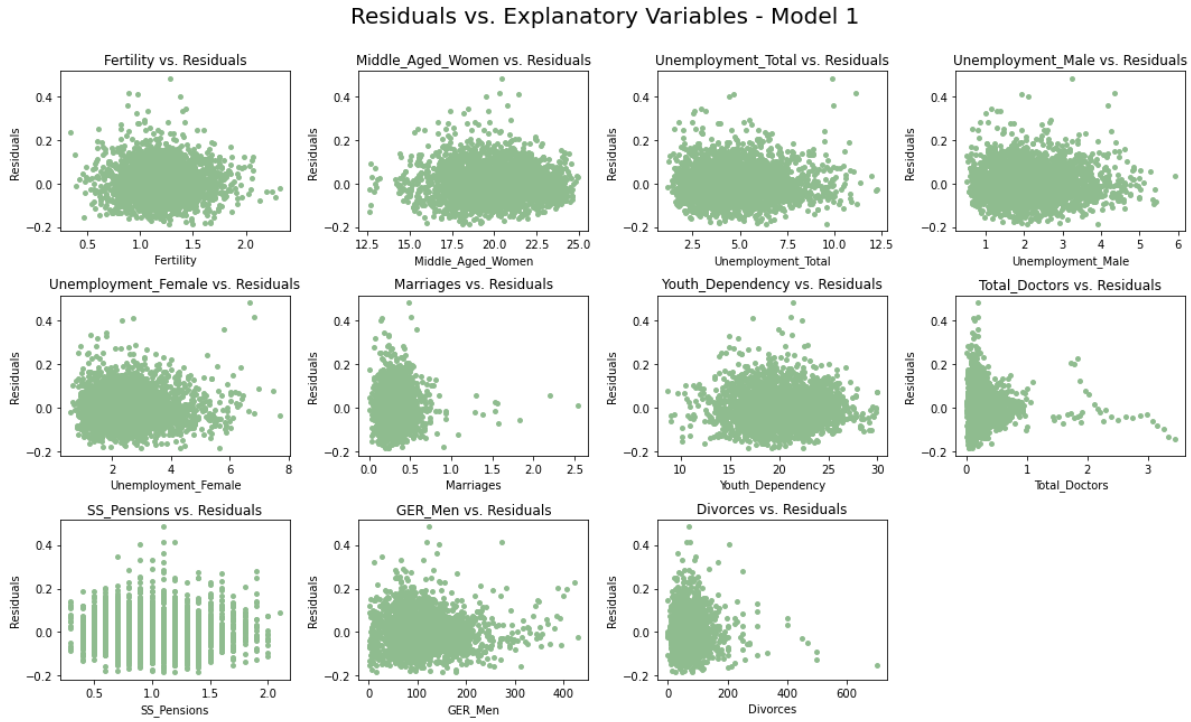


Figure 5.6 - Residuals Against Explanatory Variables (Model 1)

After analyzing the plots in Figure 5.6 there does not seem to be any evidence for the existence of a better functional form for any of the variables.

To allow better future model comparison, the adjusted R-squared was calculated for Model 1, using the following formula where N is the total number of observations and k is the total number of explanatory variables:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - k - 1}$$

Using the formula above, the adjusted R-squared for Model 1 is 0.1288.

5.7.2. Model 2 (CC)

Taking into account the results from Model 1, and still using a constant coefficients approach, a new model was estimated keeping all variables in a simple linear form and removing variables that do not seem to be relevant. This model will from now on be referred to as Model 2.

Model 2 was also estimated using cluster-robust standard errors. Its R-squared was 0.1205, lower than Model 1. This is expected since 8 variables were removed. The p-value for the overall significance test was still 0, meaning that there is really strong statistical evidence of the importance of the set of explanatory variables used in justifying the values for DVASA. Table 5.4 sums up the parameter estimates and their significancy for this model.

From Table 5.4 below, one can see that, considering a level of significance of 5%, the null hypothesis for the individual significance test cannot be rejected for some parameters. These are the ones corresponding to Middle_Aged_Women, Marriages, Youth_Dependency and GER_Men. However, if a level of significance of 10% was to be considered, the only parameters for which the null hypothesis

could not be rejected would be the ones corresponding to Middle_Aged_Women and Marriages. These values are highlighted in red on the previous table.

	Parameter	Standard Error	T-stat	P-value
Constant	0.2479	0.0734	3.3766	0.0007
Fertility	0.0307	0.0108	2.8491	0.0044
Middle_Aged_Women	-0.0042	0.0028	-1.5342	0.1251
Unemployment_Total	4.8803	2.2706	2.1493	0.0317
Unemployment_Male	-4.8805	2.2698	-2.1502	0.0316
Unemployment_Female	-4.8643	2.2716	-2.1414	0.0323
Marriages	0.0306	0.0194	1.5812	0.1139
Youth_Dependency	-0.0021	0.0011	-1.8705	0.0615
Total_Doctors	0.0451	0.0170	2.6548	0.0080
SS_Pensions	-0.0536	0.0184	-2.9112	0.0036
GER_Men	0.0001	0.000058	1.7544	0.0795
Divorces	0.0002	0.000067	2.6164	0.0089

Table 5.4 - Parameter Estimates for Model 2

In a similar way to what was done for Model 1, a graphical approach was used to test Model 2 for heteroskedasticity. The plot is shown in Figure 5.7 and, once again, no pattern of increasing or decreasing variance was detected.

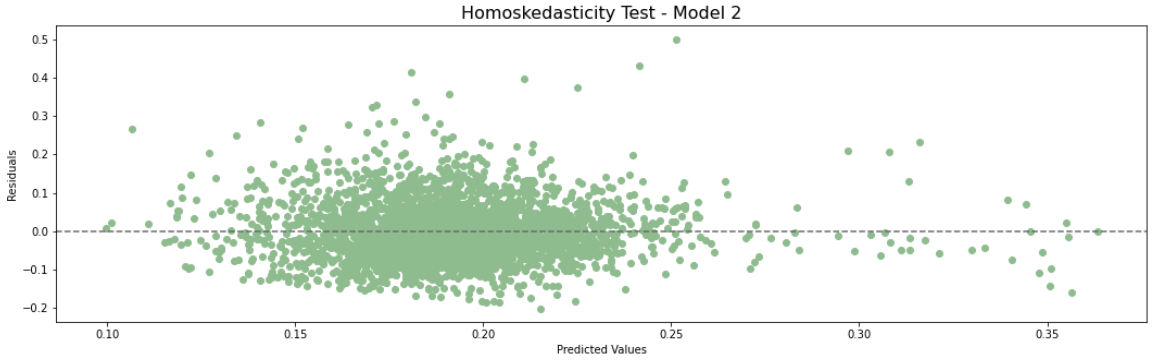


Figure 5.7 - Heteroskedasticity Test for Model 2

To have a more formal result regarding the presence of heteroskedasticity in Model 2, a Breusch-Pagan test was applied. The results for this test for this test show a really low p-value for the F-test (1.6751524272085239e-16), indicating that there is statistical evidence that at least one of the coefficients is not equal to 0, which means that there is heteroskedasticity present in the model. Consequently, the estimators are not efficient. However, since cluster-robust standard errors were used, the results of statistical tests can be evaluated. No other tests on heteroskedasticity were performed for Model 2 due to the extremely low p-value for the Breusch-Pagan test.

Similarly to what happened regarding Model 1, to test for autocorrelation among residuals the Durbin-Watson test was applied. The statistic of this test for Model 2 was 1.9072, which indicates a very small level of positive autocorrelation. Since this is a close enough value to 2, one can assume that there is no significant autocorrelation.

The value of the adjusted R-squared for Model 2 is 0.1173, lower than the value for Model 1.

5.7.3. Model 3 (CC)

According to the conclusions taken from Figure 5.4, some transformations were applied to the variables, namely converting DVASA into a natural logarithmic form. A model with all variables used for Model 1 was estimated, but this time the variables were included as log-linear relationships. This model will from now on be addressed to as Model 3. Once again, Model 3 used a constant coefficients approach with cluster-robust standard errors.

The R-squared for Model 3 was 0.1396. This means that 13.96% of the variability in ln(DVASA) is represented by this model. We can also get information on the between and within R-squared, which, similarly to the within and between standard deviations, represent the part of the differences between municipalities that is explained by the model and the part of the differences within municipalities that is explained by the model. For Model 3 the between R-squared was 0.2711 and the within R-squared was 0.0320. From these numbers, one can easily understand that a much larger part of the variance being addressed by the model is justified by differences between municipalities.

The result of the overall significance test for Model 3 was very positive, with a p-value of 0. However, according to the T-tests for individual significance, considering a significance level of 5%, there was no statistical evidence of the relevance of many variables, including the constant. The results for Model 3 can be seen on Table 5.5 below, where the p-values greater than the intended significance level (5%) are highlighted in light red.

	Parameter	Standard Error	T-stat	P-value
Constant	-0.2051	0.9792	-0.2094	0.8341
Fertility	0.1422	0.0682	2.0849	0.0372
Men65	-0.0653	0.0441	-1.4818	0.1385
Monthly_Gain	0.0002	0.0002	1.5164	0.1295
Wage_Gap	-0.0020	0.0028	-0.7318	0.4643
Middle_Aged_Women	-0.0539	0.0275	-1.9602	0.0501
Unemployment_Total	42.762	18.404	2.3235	0.0202
Unemployment_Male	-42.749	18.398	-2.3236	0.0202
Unemployment_Female	-42.685	18.410	-2.3185	0.0205
Marriages	0.1801	0.1138	1.5828	0.1136
Elderly_Dependency	0.0055	0.0080	0.6889	0.4909
Youth_Dependency	-0.0191	0.0080	-2.3778	0.0175
Female_Doctors	0.0006	0.0013	0.4505	0.6524
Total_Doctors	0.1949	0.0875	2.2265	0.0261
Mental_Health	-0.0028	0.0068	-0.4125	0.6800
SS_Pensions	-0.3451	0.1494	-2.3097	0.0210
GER	-0.0037	0.0027	-1.3367	0.1814
GER_Men	0.0029	0.0015	1.8992	0.0576
GER_Women	0.0013	0.0015	0.8358	0.4033
Divorces	0.0010	0.0004	2.4794	0.0132

Table 5.5 - Parameter Estimates for Model 3

The variables that were not statistically significant were analyzed one by one, in order to check if there were possible improvements to the functional form by converting the variables in a natural logarithmic form as well, creating log-log relationships. This was done by analyzing the plots in Figure 5.4. Men65 is included in Model 3 as a log-linear relationship in which the correspondent parameter is smaller than 0. However, it seems to resemble more the shape of a log-log relationship with the correspondent parameter being equal to -1 or below it. Thus, Men65 will be included in the next model as a log-log relationship. Monthly_Gain can also benefit from a log-log relationship with the corresponding parameter being smaller than 0. Wage_Gap does not exhibit any clear pattern and has too high of a p-value, meaning that it will be excluded from the next model. Middle_Aged_Women was included in Model 3 as a log-linear relationship with the corresponding coefficient lower than 0. However, it seems to be best described as a log-log relationship with the correspondent coefficient greater than 0. It will then be included in the next model as a log-log relationship. Marriages does not exhibit a clear pattern and will be excluded from the next model. Elderly_Dependency also seems to benefit from a log-log relationship and will be included in this way in the next model. Both Female_Doctors and Mental_Health do not exhibit clear patterns and have very high p-values, being excluded from the next model. GER and GER_Men will also be included as log-log relationships in the next model.

The adjusted R-squared for Model 3 was 0.1342.

5.7.4. Model 4 (CC)

Model 4 was, once again, estimated using a constant coefficients approach with cluster-robust standard errors. The variables used were described in the final paragraph of Chapter 5.7.3. This model presented a R-squared of 0.1395, with a within R-squared of 0.0336 and a between R-squared of 0.2687. Once again, differences between municipalities are much better explained than differences within municipalities. The p-value of the overall significance F-test was 0. The parameter estimates are summed up in Table 5.6 below. The dependent variable for this model is the natural logarithm of DVASA.

	Parameter	Standard Error	T-stat	P-value
Constant	-2.2822	2.0980	-1.0878	0.2768
Fertility	0.1445	0.0675	2.1413	0.0323
Ln_Men65	-0.6080	0.4721	-1.2878	0.1979
Ln_Monthly_Gain	0.3882	0.1351	2.8737	0.0041
Ln_Middle_Aged_Women	-0.6790	0.5012	-1.3547	0.1756
Unemployment_Total	42.802	18.588	2.3026	0.0214
Unemployment_Male	-42.794	18.581	-2.3031	0.0213
Unemployment_Female	-42.718	18.596	-2.2972	0.0217
Ln_Elderly_Dependency	0.3086	0.3816	0.80088	0.4187
Youth_Dependency	-0.00168	0.0081	-2.0704	0.0385
Total_Doctors	0.1848	0.0805	2.2963	0.0217
SS_Pensions	-0.3240	0.1416	-2.2876	0.0222
Ln_GER	0.0064	0.0659	0.0974	0.9224
Ln_GER_Men	0.0792	0.0538	1.4707	0.1415
GER_Women	-0.0003	0.0004	-0.775	0.4384

Divorces	0.0006	0.0004	1.6525	0.0985

Table 5.6 - Parameter Estimates for Model 4

One can see that for most of the transformed variables there is no statistical evidence of their significance to the model. This can be due to the lower variance inside municipalities, as the variables are not able to explain differences between observations for different years inside the same municipality. P-values that do not allow for the rejection of the null hypothesis according to a 10% significance level are highlighted in red. The adjusted R-squared for Model 4 was 0.1353.

The Breusch-Pagan test was applied to the residuals of this model, resulting in a p-value of 1.4599286323148384e-36, which indicates clear presence of heteroskedasticity. Since the classical assumptions do not apply in all four models estimated with a constant coefficients approach, applying Fixed or Random Effects models to this dataset seems to be a better option.

5.7.5. Model 5 (FE – Fixed Effects)

Since the models previously built had higher values for the between R-squared than for the within R-squared, one can conclude that differences between municipalities are being explained more efficiently than differences within municipalities (between different years). Keeping this in mind, a fixed-effects model with different constants for each year was built. This consisted of estimating a model while including dummies for each year with the purpose of indicating whether an observation refers to that year. The year 2009 did not have a correspondent dummy to avoid multicollinearity issues. Thus, the regular constant of the model represents the constant for 2009 and the coefficients of the remaining dummies represent the differences between constants for respective years and the constant for 2009. All 19 explanatory variables were included in a simple linear form. The model can be hypothesized as follows:

$$DVASA_{it} = \alpha_0 + \alpha_1 D2010_i + \alpha_2 D2011_i + \alpha_3 D2012_i + \alpha_4 D2013_i + \alpha_5 D2014_i + \alpha_6 D2015_i + \alpha_7 D2016_i + \alpha_8 D2017_i + \alpha_9 D2018_i + \alpha_{10} D2019_i + \beta_1 Fertility_i + \beta_2 Men65_i + \cdots + \beta_{19} Divorces_i + u_i$$

The estimated model showed an overall R-squared of 0.1664 with a within R-squared of 0.0462 and a between R-squared of 0.2957. The p-value for the overall significance F-test was 0, showing once again the importance of the variables taken into consideration for explaining changes in DVASA. The adjusted R-squared for this model was 0.1584. The parameter estimates for this model, from now on called Model 5, can be seen on Table 5.7 below. All values for the p-value of the individual significance test which, considering a significance level of 10%, fail to reject the null hypothesis are highlighted in red.

	Parameter	Standard Error	T-stat	P-value
Constant	0.2658	0.1663	1.5977	0.1102
Fertility	0.0201	0.0105	1.9219	0.0547
Men65	-0.0073	0.0062	-1.1846	0.2363
Monthly_Gain	-0.0000009	0.000026	-0.0345	0.9725
Wage_Gap	-0.0011	0.0005	-2.3469	0.0190
Middle_Aged_Women	-0.0043	0.0047	-0.9248	0.3551
Unemployment_Total	3.8365	2.2843	1.6795	0.0932
Unemployment_Male	-3.8273	2.2832	-1.6763	0.0938

Unemployment_Female	-3.8227	2.2852	-1.6728	0.0945
Marriages	0.0382	0.0183	2.0895	0.0367
Elderly_Dependency	0.0007	0.0011	0.6916	0.4892
Youth_Dependency	0.0008	0.0016	0.4869	0.6264
Female_Doctors	-0.0002	0.0002	-0.7914	0.4288
Total_Doctors	0.0377	0.0805	2.2963	0.0217
Mental_Health	-0.0004	0.0010	-0.3960	0.6922
SS_Pensions	-0.0083	0.0244	-0.3418	0.7325
GER	-0.0009	0.0005	-1.9691	0.0490
GER_Men	0.0006	0.0003	2.0253	0.0429
GER_Women	0.0005	0.0002	2.1303	0.0332
Divorces	0.0002	0.00006	2.7776	0.0055
D2010	0.0185	0.0045	4.1338	0.0000
D2011	0.0105	0.0060	1.7370	0.0825
D2012	-0.0008	0.0075	-0.1040	0.9172
D2013	0.0049	0.0077	0.6354	0.5252
D2014	0.0161	0.0078	2.0486	0.0406
D2015	0.0212	0.0081	2.6101	0.0091
D2016	0.0269	0.0088	3.0475	0.0023
D2017	0.0377	0.0095	3.9588	0.0001
D2018	0.0507	0.0105	4.8258	0.0000
D2019	0.0711	0.0119	5.9929	0.0000

Table 5.7 - Parameter Estimates for Model 5

One can see from the analysis of Table 5.7 above that there are some very high p-values for the individual significance T-test on some variables, namely Monthly_Gain and D2012. This means that the effects of these variables are probably not relevant for the model, and they should be removed. The fact that the p-value for the individual significance test on the coefficient of D2012 is so high means that there is probably no significant difference between the constant for 2009 and the constant for 2012.

The Breusch-Pagan test was applied to the residuals of Model 5, resulting in a p-value of 3.818051373052344e-16, which indicates clear presence of heteroskedasticity. Since the parameters were calculated using cluster-robust standard errors, the statistical inference is valid. However, the estimators are not efficient.

To test for autocorrelation among residuals the Durbin-Watson test was applied. The statistic of this test for Model 5 was 1.9245, which indicates a very small level of positive autocorrelation. Since this is a close enough value to 2, one can assume that there is no significant autocorrelation.

5.7.6. Model 6 (FE)

Since so many of the variables used for Model 5 did not show statistical evidence of having a real impact on DVASA occurrences, some of them were removed and a new model (Model 6) was estimated following the same logic followed for Model 5. However, this time, the dummy for 2012 was not included. The other explanatory variables removed were Monthly_Gain, SS_Pensions, Mental_Health and Youth_Dependency. The R-squared value for Model 6 was 0.1657, which is surprisingly not a huge

drop from the R-squared of Model 5 considering that it has 5 variables less. The value for the within R-squared was 0.0463 and for the between R-squared was 0.2941. The parameter estimates for this model can be seen on Table 5.8 below. All values for the p-value of the individual significance test which, considering a significance level of 10%, fail to reject the null hypothesis are highlighted in red.

	Parameter	Standard Error	T-stat	P-value
Constant	0.2769	0.1344	2.0600	0.0395
Fertility	0.0260	0.0110	2.3724	0.0177
Men65	-0.0079	0.0057	-1.4005	0.1615
Wage_Gap	-0.0011	0.0003	-3.1171	0.0018
Middle_Aged_Women	-0.0042	0.0044	-0.9645	0.3349
Unemployment_Total	3.6990	2.2755	1.6256	0.1041
Unemployment_Male	-3.6901	2.2744	-1.6224	0.1048
Unemployment_Female	-3.6857	2.2762	-1.6192	0.1055
Marriages	0.0397	0.0181	2.1941	0.0283
Elderly_Dependency	0.0006	0.0010	0.6071	0.5438
Female_Doctors	-0.0001	0.0002	-0.7346	0.4627
Total_Doctors	0.0373	0.0147	2.5363	0.0113
GER	-0.0009	0.0005	-1.9093	0.0563
GER_Men	0.0005	0.0003	1.9730	0.0486
GER_Women	0.0004	0.0002	2.0532	0.0401
Divorces	0.0002	0.00007	2.6535	0.008
D2010	0.0188	0.0037	5.1622	0.0000
D2011	0.0106	0.0039	2.6932	0.0071
D2013	0.0054	0.0047	1.1585	0.2468
D2014	0.0164	0.0051	3.1901	0.0014
D2015	0.0208	0.0051	4.0788	0.0000
D2016	0.0261	0.0056	4.6801	0.0000
D2017	0.0369	0.0058	6.3273	0.0000
D2018	0.0495	0.0064	7.7633	0.0000
D2019	0.0703	0.0074	9.4873	0.0000

Table 5.8 - Parameter Estimates for Model 6

By analyzing Table 5.8, one can see that the differences in the constant by year are mostly very relevant. Almost all coefficients for the yearly dummies show very low p-values for their individual significance test, except for the one corresponding to 2013, which does not show statistical evidence of being relevant for the model. Some explanatory variables still show very high associated p-values for individual significance tests, such as Elderly_Dependency, Female_Doctors and Middle_Aged_Women, and should probably be removed from the model.

The Breusch-Pagan test was applied to the residuals of Model 6, resulting in a p-value of 6.853637246359593e-16, which indicates clear presence of heteroskedasticity. Since the parameters were calculated using cluster-robust standard errors, the statistical inference is valid. However, the estimators are not efficient.

To test for autocorrelation among residuals the Durbin-Watson test was applied. The statistic of this test for Model 6 was 1.9259, which indicates a very small level of positive autocorrelation. Since this is a close enough value to 2, one can assume that there is no significant autocorrelation.

5.7.7. Model 7 (FE)

Taking into account the results obtained for Model 6, a new model was estimated using a fixed effects approach. To estimate Model 7, variables that were not showing evidence of relevance were removed – Men65, Middle_Aged_Women, Elderly_Dependency, Female_Doctors and D2013. The R-squared for Model 7 was 0.1607, with a between R-squared of 0.2918 and a within R-squared of 0.0389. The overall significance F-test for this model showed a p-value of closely 0. The parameter estimates for this model can be seen on Table 5.9 below. All variables showed statistical evidence of having coefficients different than 0, considering a 10% significance level.

	Parameter	Standard Error	T-stat	P-value
Constant	0.1192	0.0301	3.9591	0.0001
Fertility	0.0311	0.0105	2.9527	0.0032
Wage_Gap	-0.0011	0.0003	-3.3348	0.0009
Unemployment_Total	3.9272	2.2769	1.7248	0.0847
Unemployment_Male	-3.9180	2.2759	-1.7215	0.0853
Unemployment_Female	-3.9120	2.2776	-1.7176	0.0860
Marriages	0.0412	0.0182	2.2622	0.0238
Total_Doctors	0.0367	0.0140	2.6225	0.0088
GER	-0.0008	0.0005	-1.7297	0.0838
GER_Men	0.0005	0.0003	1.8274	0.0677
GER_Women	0.0004	0.0002	1.8488	0.0646
Divorces	0.0002	0.00007	2.6260	0.0087
D2010	0.0176	0.0038	4.5983	0.0000
D2011	0.0094	0.0040	2.3738	0.0177
D2014	0.0134	0.0044	3.0525	0.0023
D2015	0.0175	0.0044	4.0281	0.0001
D2016	0.0224	0.0048	4.7141	0.0000
D2017	0.0340	0.0052	6.4950	0.0000
D2018	0.0470	0.0060	7.8414	0.0000
D2019	0.0685	0.0069	9.8940	0.0000

Table 5.9 - Parameter Estimates for Model 7

The Breusch-Pagan test was applied to the residuals of Model 7, resulting in a p-value of 3.910166886806135e-06, which indicates clear presence of heteroskedasticity. Since the parameters were calculated using cluster-robust standard errors, the statistical inference is valid. However, the estimators are not efficient.

To test for autocorrelation among residuals the Durbin-Watson test was applied. The statistic of this test for Model 7 was 1.9443, which indicates a very small level of positive autocorrelation. Since this is a close enough value to 2, one can assume that there is no significant autocorrelation.

Even if having all variables being significant, the interpretation of this model is not straightforward due to the fact that variables regarding unemployment by gender and total unemployment are included in the same model. The same is true for the variables regarding education. Keeping this in mind, the variables Unemployment_Total and GER will be removed from the next model.

5.7.8. Model 8 (FE)

Model 8 was estimated to improve interpretability of results. It is very similar to Model 7 but does not include Unemployment_Total and GER. The parameter estimates for this model can be seen on Table 5.10 below. All values for the p-value of the individual significance test which, considering a significance level of 10%, fail to reject the null hypothesis are highlighted in red.

	Parameter	Standard Error	T-stat	P-value
Constant	0.1176	0.0302	3.8948	0.0001
Fertility	0.0312	0.0106	2.9439	0.0033
Wage_Gap	-0.0011	0.0003	-3.2289	0.0013
Unemployment_Male	0.0093	0.0069	1.3568	0.1749
Unemployment_Female	0.0150	0.0071	2.1051	0.0354
Marriages	0.0406	0.0182	2.2274	0.0260
Total_Doctors	0.0364	0.0139	2.6099	0.0091
GER_Men	0.0001	0.0001	1.0528	0.2925
GER_Women	0.000008	0.00009	0.0875	0.9303
Divorces	0.0002	0.00007	2.5984	0.0094
D2010	0.0174	0.0038	4.5830	0.0000
D2011	0.0097	0.0040	2.4577	0.0140
D2014	0.0134	0.0044	3.0497	0.0023
D2015	0.0173	0.0044	3.9516	0.0001
D2016	0.0218	0.0048	4.5517	0.0000
D2017	0.0338	0.0053	6.3528	0.0000
D2018	0.0467	0,0060	7.7715	0.0000
D2019	0.0686	0.0069	9.8865	0.0000

Table 5.10 - Parameter Estimates for Model 8

It becomes clear from the analysis of Table 5.10 that the variables related to unemployment by gender and education by gender lose significance if the total variables are removed. Considering this, an alternative model was then estimated including the total variables for unemployment and education and excluding the ones by gender.

5.7.9. Model 9 (FE)

The parameter estimates for Model 9 can be seen on Table 5.11 below. The R-squared for Model 9 was 0.1566, with a between R-squared of 0.2877 and a within R-squared of 0.0349. The overall significance F-test for this model showed a p-value of 0. All variables showed statistical evidence of having coefficients different than 0, considering a 5% significance level. This will be the final model for this study.

	Parameter	Standard Error	T-stat	P-value
Constant	0.1188	0.0302	3.9396	0.0001
Fertility	0.0303	0.0109	2.7727	0.0056
Wage_Gap	-0.0011	0.0003	-3.2945	0.0010
Unemployment_Total	0.0125	0.0020	6.2752	0.0000
Marriages	0.0414	0.0185	2.2429	0.0250
Total_Doctors	0.0364	0.0138	2.6325	0.0085
GER	0.0001	0.00006	1.9742	0.0485
Divorces	0.0002	0.00007	2.5852	0.0098
D2010	0.0180	0.0036	4.9379	0.0000
D2011	0.0106	0.0042	2.5163	0.0119
D2014	0.0133	0.0043	3.0797	0.0021
D2015	0.0175	0.0043	4.0435	0.0001
D2016	0.0222	0.0047	4.7126	0.0000
D2017	0.0345	0.0055	6.3351	0.0000
D2018	0.0477	0.0063	7.5241	0.0000
D2019	0.0697	0.0072	9.7471	0.0000

Table 5.11 - Parameter Estimates for Model 9

The Breusch-Pagan test was applied to the residuals of Model 9, resulting in a p-value of 0.00196, which indicates clear presence of heteroskedasticity. Since the parameters were calculated using cluster-robust standard errors, the statistical inference is valid. However, the estimators are not efficient.

To test for autocorrelation among residuals the Durbin-Watson test was applied. The statistic of this test for Model 9 was 1.9416, which indicates a very small level of positive autocorrelation. Since this is a close enough value to 2, one can assume that there is no significant autocorrelation.

6. RESULTS AND DISCUSSION

Nine models were created in the scope of this study. Four of them used the constant coefficients approach and the remaining five used the fixed effects approach. The objective was to build a model that is substantiated by statistical evidence. For this purpose, results of overall significance tests and individual significance tests must by analyzed.

The R^2 or coefficient of determination of a regression is a measure of the proportion of variability in the dependent variable that is justified by changes in the explanatory variables. However, the R^2 may not always be the best measure to compare models as it always becomes larger when adding new variables, even if the variables added are completely irrelevant. Keeping this in mind, the R^2 can only be used to compare models with the exact same number of explanatory variables. When the models have a different number of explanatory variables, the adjusted R^2 can be used as an alternative.

Table 6.1 below shows a summary of the measures calculated for the models created to facilitate comparison between them. In this table, one can see the dependent variable used for each model, the number of explanatory variables included in each model, the values for the overall R-squared, between R-squared, within R-squared and adjusted R-squared for each model, the number of the variables included as explanatory variables that had a p-value lower than 0.1 and were, therefore, considered relevant based on a 10% significance level, the p-value for the overall significance F-test, the p-value for the Breusch-Pagan test and the value for the Durbin-Watson statistic.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Dependent Variable	DVASA	DVASA	Ln(DVASA)	Ln(DVASA)	DVASA	DVASA	DVASA	DVASA	DVASA
Nº Explanatory Variables	19	11	19	15	29	24	19	17	15
Nº Observations	3,058	3,058	3,058	3,058	3,058	3,058	3,058	3,058	3,058
R-squared	0.1342	0.1205	0.1396	0.1395	0.1664	0.1657	0.1607	0.1577	0.1566
Between R-squared	0.2592	0.2460	0.2711	0.2687	0.2957	0.2941	0.2918	0.2882	0.2877
Within R-squared	0.0181	0.0040	0.0320	0.0336	0.0462	0.0463	0.0389	0.0365	0.0349
Adjusted R-squared	0.1288	0.1173	0.1342	0.1353	0.1584	0.1591	0.1556	0.1530	0.1524
Nº Relevant Variables (10%)	10	9	10	9	19	16	19	14	15
Overall Significance	0	0	0	0	0	0	0	0	0
Breusch-Pagan	≈0	≈0	≈0	≈0	≈0	≈0	≈0	≈0	0.00196
Durbin-Watson	1.8941	1.9072	1.9151	1.9068	1.9245	1.9259	1.9443	1.9396	1.9416

Table 6.1 - Model Comparison

Model 1, the first one created, included all retrieved possible explanatory variables in a simple linear form. It showed promising results concerning the value for the between R-squared. However, almost

half of the variables could not be considered relevant and the value for the within R-squared was extremely low, exposing from the start the problem of low variance of the data within municipalities. To try and achieve a more statistically relevant model, Model 2 was created, removing eight of the non-relevant variables exposed in Model 1 while keeping all relationships in a simple linear form. The remaining two non-relevant variables from Model 1 were kept in Model 2 to see if their behavior changed when included in a more restricted set of explanatory variables. However, their corresponding individual significance tests continued to show them non-relevant. As the value for the R-squared was very low, Model 3 was created using different functional forms (log-linear) to check for improvements in explainability. This model showed promising results but many variables had high p-values for their individual significance test, causing the model to be not very reliable. A further attempt with the natural logarithm of DVASA as the dependent variable was made in Model 4, but this time including some of the variables in a log-log form and removing the ones that did not show apparent patterns when plotted against DVASA. Once again, this model showed promising results but many variables had high p-values for their individual significance tests. As all the previous models were showing clear evidence of heteroskedasticity, a change in the model approach was considered, resulting in Model 5, estimated using fixed effects for the time component. In this model, differences between constants for different years were included using dummy variables for all years except 2009, to avoid multicollinearity. From Model 5 to Model 9, different combinations of variables were tried in order to optimize interpretability and reliance of the model.

All of the estimated models ended up showing evidence of the presence of heteroskedasticity. This means that the estimators are not efficient and that, if cluster-robust standard errors had not been used in all estimations, statistical inference would not be valid.

As already mentioned, valid statistical evidence is the main criteria when it comes to choosing the final model. Thus, all variables of that model must be significant and the model itself must pass the overall significance test. The only two models that meet these criteria are Model 7 and Model 9. Since the number of variables of these two models is different, the R-squared must not be used as a comparison measure. Instead, the comparison measure to be used is the adjusted R-squared. However, even if Model 7 shows a higher value for this measure, Model 9 has better interpretability, making it a serious contestant for final model. A choice was made to value interpretability in this case as the difference between R-squared values is not much significant. That being said, Model 9 was chosen as the final model.

7. CONCLUSIONS

As stated in the Introduction, the present dissertation proposed to answer the following questions:

- How did the number of domestic violence occurrences in Portugal evolve between 2009 and 2019?
- How well can panel data regression explain this evolution?
- What are the main causes of domestic violence?
- How does each explanatory variable affect the number of domestic violence occurrences?

In order to achieve the proposed goal, several models were estimated using the number of domestic violence occurrences against spouse or analogous or variations of this as the dependent variable. First of all, it is possible to conclude that explaining the evolution of domestic violence occurrences is a difficult process as variables on the individual level cannot be included to explain occurrences in a municipality. Due to this, the models estimated ended up justifying approximately 15 or 16% of the total variance in the dependent variable. The model selected as final can be written as following:

$$\begin{aligned} DVASA_i = {} & 0.1188 + 0.0180D2010_i + 0.0106D2011_i + 0.0133D2014_i + 0.0175D2015_i + 0.0222D2016_i \\ & + 0.0345D2017_i + 0.0477D2018_i + 0.0697D2019_i + 0.0303Fertility_i \\ & + 0.0002Divorces_i + 0.0001GER_i + 0.0414Marriages_i - 0.0011Wage\_Gap_i \\ & + 0.0125Unemployment\_Total_i + 0.0364Total\_Doctors_i + e_i \end{aligned}$$

This means that in 2009, 2012 and 2013, if all the other variables are set to 0 and holding everything else constant, on average, the number of DVASA occurrences registered by police authorities per 100 inhabitants was 0.1188. In 2010, it is estimated that municipalities had, on average and holding all other factors constant an extra 0,018 DVASA occurrences registered by police authorities when compared to 2009, 2012 and 2013. According to the same rationale, for 2011, 2014, 2015, 2016, 2017, 2018 and 2019, it is expected that municipalities, holding all other factors constant, have an extra 0.0106, 0.0133, 0.0175, 0.0222, 0.0345, 0.0477 and 0.0697, respectively, DVASA occurrences registered by police authorities for 100 inhabitants when compared to 2009, 2012 and 2013.

Fertility represents the average number of children born to each woman in fertile age. The coefficient for this variable is 0.0303, which means that, holding all other factors constant, it is expected that an increase of one child *per* woman in fertile age causes an increase of 0.0303 DVASA occurrences *per* 100 inhabitants in that same municipality. This corroborates the idea stated in (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001) that claims that women with more children are more likely to become victims of domestic violence.

The variable Divorces represents the number of divorces per 100 marriages. Similarly to Fertility, this variable has also proven to be a risk factor. As the number of divorces increases, the number of DVASA occurrences is also expected to increase, which can be seen in the positive coefficient associated to this variable in the model. One can say that, if the number of divorces by 100 marriages increases by 10, holding all other factors constant, it is expected that the number of DVASA occurrences per 100 inhabitants increases by 0.002. This corroborates the theory stated in (WHO - World Health Organization, 2010), affirming that divorces cannot only be a consequence of domestic violence, but also a cause. People who are separated or divorced tend to be more vulnerable, increasing their probability of becoming victims of domestic violence.

GER represents the percentage of resident population with normal age for attending high school that is actually attending high school. The coefficient for this variable being positive was an astonishing find. It goes against all preconceived ideas and all ideas exposed in the literature review. It shows that as the percentage represented in GER increases, it is expected that so does the number of DVASA occurrences per 100 inhabitants. This variable was retrieved due to lack of data regarding the education level of the population by municipality, so it might be interesting to try the same regression using a more appropriate variable, if made available.

The variable Marriages represents the number of new marriages per 100 inhabitants and has a positive coefficient associated, which makes it a risk factor. This was expected as most of domestic violence victims are married. However, this was a controversial variable as it should be added in a way that reflects past marriages. Nevertheless, municipalities with more marriages may also be more conservative, which adds a new hypothesis on the table. From the model one can see that it is expected that for each new marriage by 100 inhabitants, the number of DVASA occurrences by 100 inhabitants increases by 0.0414, holding all other factors constant.

Wage_Gap represents the percentage of men's monthly pay that women receive on average. The coefficient for this variable is negative, which meets the theory of exposure reduction, explained in (Aizer, 2010) which states that, as the wage gap decreases, the labor force participation of women increases and, consequently, domestic violence against them declines because women spend less time with violent partners. According to the model built, for each percentual point that women salary gains when faced with men monthly gain, it is expected that the number of DVASA occurrences by 100 inhabitants decreases by 0.0011, holding all other factors constant.

Unemployment_Total represents the number of people enrolled in employment and vocational training centers per 100 inhabitants. As expected, the coefficient for this variable is positive, making unemployment a risk factor. According to the model, it is expected that for each 10 additional persons enrolled in employment and vocational training centers, the number of DVASA occurrences per 100 inhabitants increases by 0.0125.

Finally, Total_Doctors was added to the dataset to try and fight underreporting, as doctors are the professionals who deal directly with the consequences of domestic violence and can report occurrences to the authorities. This variable represents the number of doctors per 100 inhabitants according to the Doctors' Professional Order. It has a positive associated coefficient, as expected. According to the model built, it is expected that each new doctor per 100 inhabitants leads to an increase of reported DVASA occurrences by 100 inhabitants of 0.0364, holding all other factors constant.

8. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Explaining a good proportion of the total variance encapsulated in the dependent variable was revealed to be a hard task. When choosing to study domestic violence occurrences using a social structural theory, it is almost inevitable that explanatory variables are most of the times very static. A lot of macro-level variables tend to evolve very slowly, causing a problem of low variance in short time series. This is what happened with data gathered for each municipality and this is the reason why, for every model created, much higher values were presented for the between R-squared than for the within R-squared. Nevertheless, the main goal of a model like the ones built in the present document is to allow authorities to act on the problem and, for that matter, differences on a spatial level are more important than on a temporal level. Nevertheless, it is my opinion that studies focused on the individual-level should be done to find more prominent causes of domestic violence.

Furthermore, it is impossible to know for sure whether the variance in the dependent variable reflects reality. As mentioned before, domestic violence is a crime that commonly takes place in the privacy of a home and, for that reason, many cases may depend on self-report. This means that according to (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001), the number of occurrences registered may suffer from underreporting. Explanatory variables also have their own issues. Due to unavailability of data, some explanatory variables are not exactly those originally thought for the theoretical model. It is the case of all variables related to unemployment, variables related to education, among others. The variables that ended up being used are as close as possible to the original idea, but they might have led to a not so ideal set of independent variables in the final model.

Finally, the only two panel data analysis approaches used were the constant coefficients and the fixed effects. There was an attempt to apply a random effects approach, but it did not show very promising results and, considering the scope of this work, was not pursued. However, I believe that it may be possible to build interesting models on this subject using random effects. Another option would be to try different slopes for different years by creating interactions between the dummy variables already created and other explanatory variables.

9. BIBLIOGRAPHY

Ackerson, L., Kawachi, I., Barbeau, E., & Subramanian, S. (2008). Effects of Individual and Proximate Educational Context on Intimate Partner Violence: A Population-Based Study of Women in India. *American Journal of Public Health*, 507-514.

Aizer, A. (2010). The Gender Wage Gap and Domestic Violence. *American Economic Review*, 1847-1859.

Amirthalingam, K. (2005). Women's Rights, International Norms, and Domestic Violence:. *Human Rights Quarterly,*, 683-708.

Anderberg, D., Rainer, H., Wadsworth, J., & Wilson, T. (2015). Unemployment and Domestic Violence: Theory and Evidence. *The Economic Journal*, 1947-1979.

APAV - Associação Portuguesa de Apoio à Vítima. (2018). *Homens Vítimas de Violência Doméstica 2013 - 2017.* Lisboa: Associação Portuguesa de Apoio à Vítima.

APAV - Associação Portuguesa de Apoio à Vítima. (2018). *Vítimas de Violência Doméstica 2013 - 2017.* Lisboa: Associação Portuguesa de Apoio à Vítima.

APAV - Associação Portuguesa de Apoio à Vítima. (2020). *Vítimas de Homicído Relatório APAV 2019.* Lisboa: Associação Portuguesa de Apoio à Vítima.

Barber, C. (2008). Domestic Violence Against Men. *Nursing Standard*, 35-39.

Bowlus, A., & Seitz, S. (2006). Domestic Violence, Employment and Divorce. *International Economic Review*, 1113-1149.

Brasil, E., Alves, F., & Soares, S. (2018). *Dados 2017.* Almada: OMA - Observatório de Mulheres Assassinadas da UMAR.

Brasil, E., Alves, F., & Soares, S. (2019). *Dados 2018.* Almada: OMA - Observatório de Mulheres Assassinadas da UMAR.

Campbell, J. (2002). Health Consequences of Intimate Partner Violence. *The Lancet*, 1331-1336.

Cann, K., Withnell, S., Shakespeare, J., Doll, H., & Thomas, J. (2001). Domestic Violence: A Comparative Survey of Levels of Detection, Knowledge and Attitudes in Healthcare Workers. *Public Health*, 89-95.

Devries, M., Mak, T., Garcia-Moreno, C., Petzold, M., Child, C., Falder, G., . . . Watts, H. (2013). The Global Prevalence of Intimate Partner Violence Against Women. *Science*, 1527-1528.

Ellsberg, M., Heise, L., Peña, R., Agurto, S., & Winkvist, A. (2001). Researching Domestic Violence Against Women: Methodological and Ethical Considerations. *Studies in Family Planning*.

Hill, R. C., Griffiths, W. E., & Lim, G. C. (2012). *Principles of Econometrics.* John Wiley and Sons.

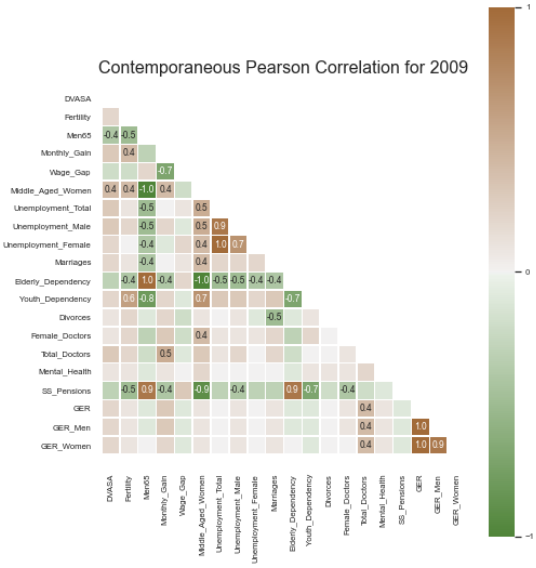
Soares, S., Branco, E., & Alves, F. (2020). *Relatório Anual 2019.* Almada: OMA - Observatório de Mulheres Assassinadas da UMAR.

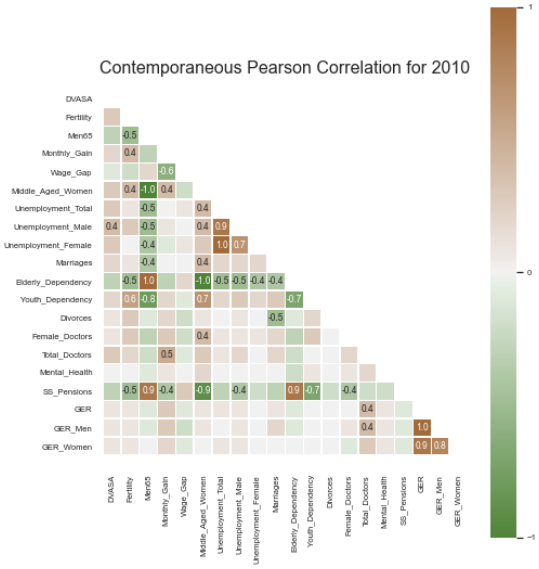
WHO - World Health Organization. (2010). *Preventing Intimate Partner and Sexual Violence Against Women: Taking Action and Generating Evidence.* Geneva: World Health Organization.

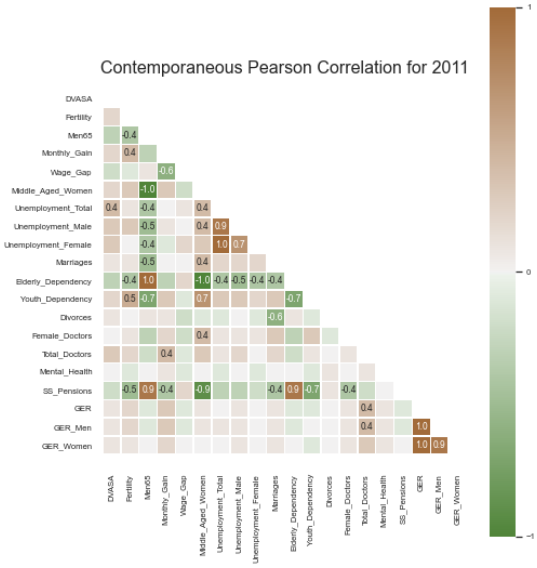
Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach.* Michigan: South-Western.

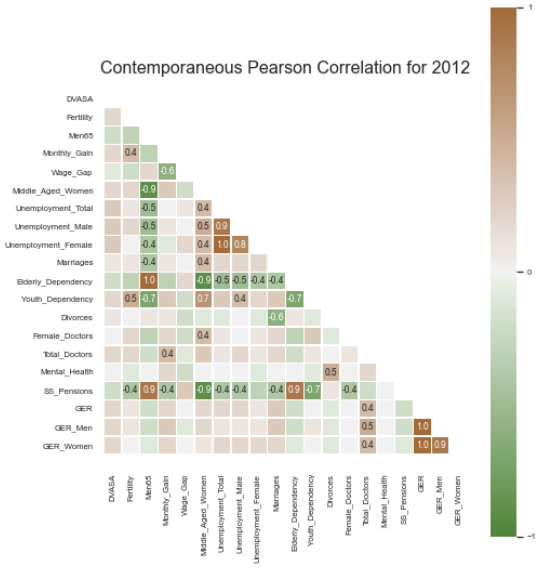
10. ANNEXES

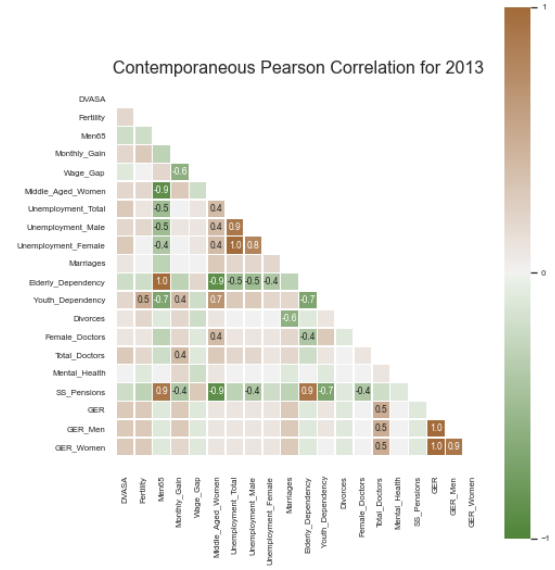
Annex I. Contemporaneous Correlation Matrixes

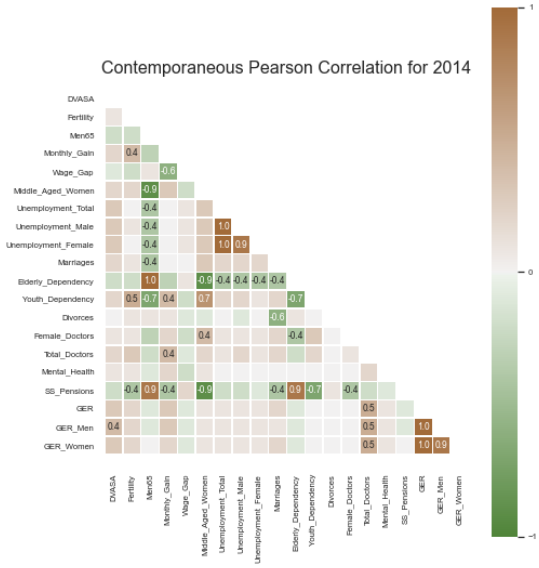


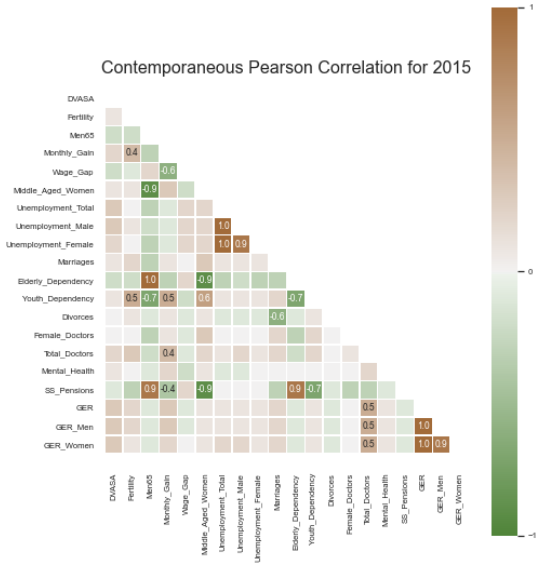


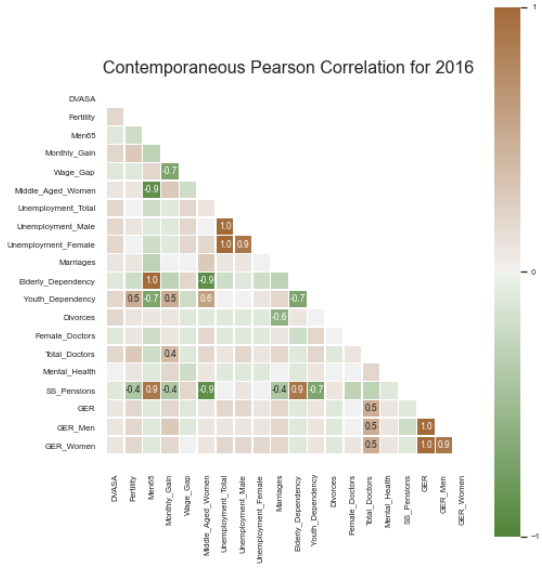


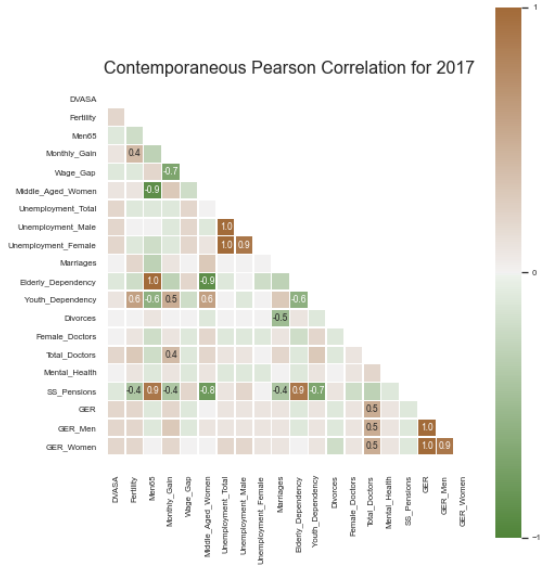


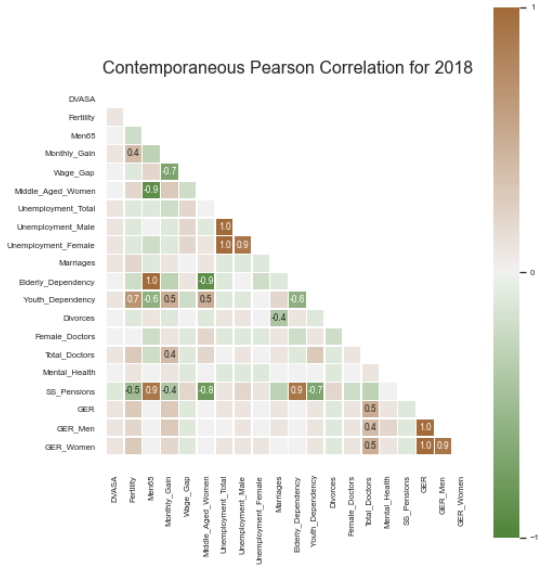


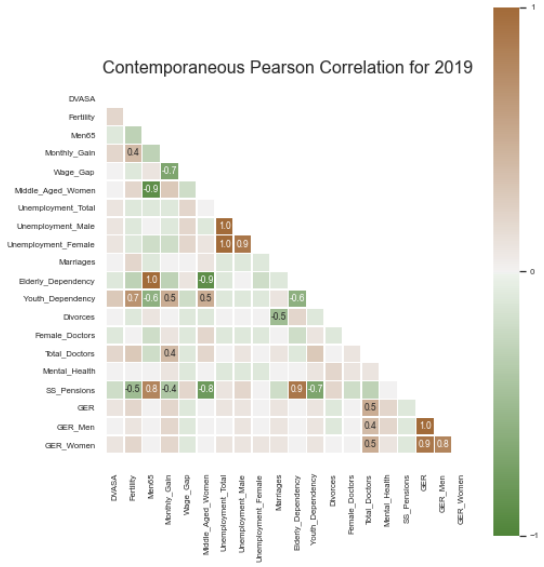



Contemporaneous Pearson Correlation for 2015
DVASA
Fertility
Men65
Monthly_Gain
Wage_Gap
Middle_Aged_Women
Unemployment_Total
Unemployment_Male
Unemployment_Female
Marriages
Elderly_Dependency
Youth_Dependency
Divorces
Female_Doctors
Total_Doctors
Mental_Health
SS_Pensions
GER
GER_Men
GER_Women

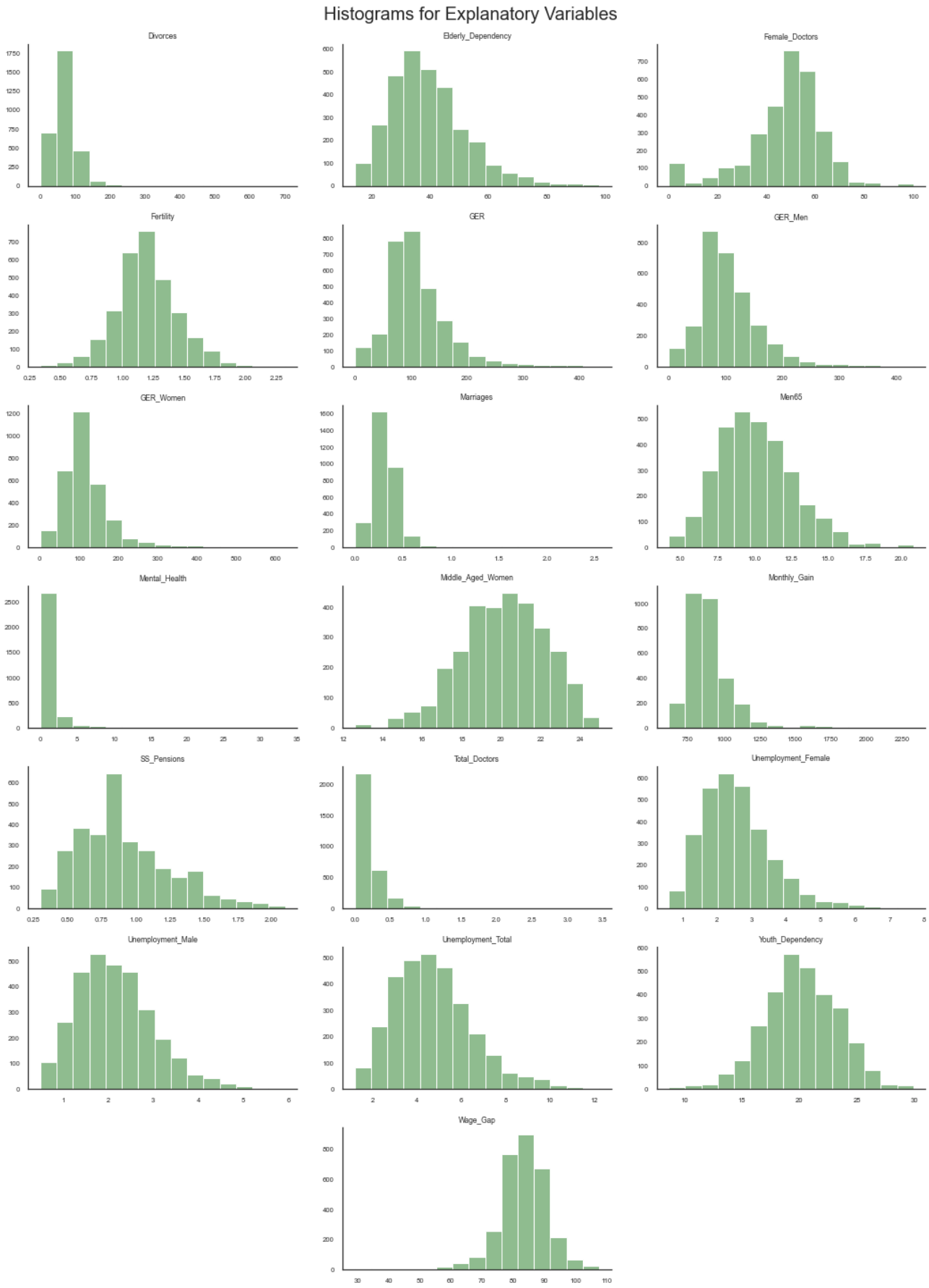

Contemporaneous Pearson Correlation for 2016
DVASA
Fertility
Men65
Monthly_Gain
Wage_Gap
Middle_Aged_Women
Unemployment_Total
Unemployment_Male
Unemployment_Female
Marriages
Elderly_Dependency
Youth_Dependency
Divorces
Female_Doctors
Total_Doctors
Mental_Health
SS_Pensions
GER
GER_Men
GER_Women


Contemporaneous Pearson Correlation for 2017
DVASA
Fertility
Men65
Monthly_Gain
Wage_Gap
Middle_Aged_Women
Unemployment_Total
Unemployment_Male
Unemployment_Female
Marriages
Elderly_Dependency
Youth_Dependency
Divorces
Female_Doctors
Total_Doctors
Mental_Health
SS_Pensions
GER
GER_Men
GER_Women

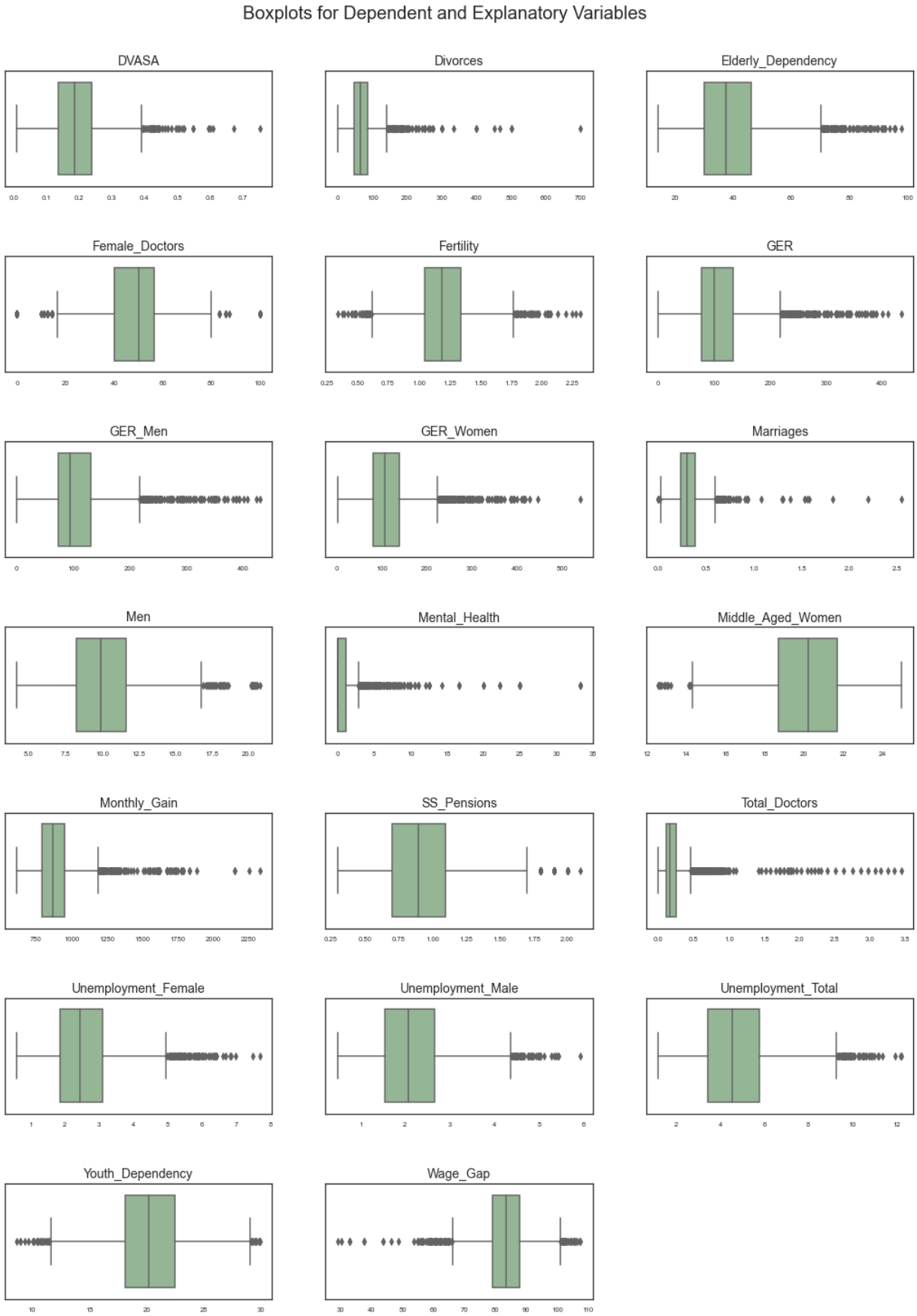

Contemporaneous Pearson Correlation for 2018
DVASA
Fertility
Men65
Monthly_Gain
Wage_Gap
Middle_Aged_Women
Unemployment_Total
Unemployment_Male
Unemployment_Female
Marriages
Elderly_Dependency
Youth_Dependency
Divorces
Female_Doctors
Total_Doctors
Mental_Health
SS_Pensions
GER
GER_Men
GER_Women


Contemporaneous Pearson Correlation for 2019
DVASA
Fertility
Men65
Monthly_Gain
Wage_Gap
Middle_Aged_Women
Unemployment_Total
Unemployment_Male
Unemployment_Female
Marriages
Elderly_Dependency
Youth_Dependency
Divorces
Female_Doctors
Total_Doctors
Mental_Health
SS_Pensions
GER
GER_Men
GER_Women

Annex II – Histograms for Explanatory Variables



Annex III – Boxplots for Dependent and Explanatory Variables



Annex IV – Evolution and First Differences for All Variables