



Analyzing Incomplete Voting Data: The Case of European Social Survey Round 10 in Greece

By

Melini Yfanti-Kostopoulou

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece

September 2024







ΑΝΑΛΥΟΝΤΑΣ ΕΛΛΙΠΗ ΔΕΔΟΜΕΝΑ ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΨΗΦΟ: Η ΠΕΡΙΠΤΩΣΗ ΤΟΥ ΚΥΚΛΟΥ 10 ΤΗΣ ΕΥΡΩΠΑΪΚΗΣ ΚΟΙΝΩΝΙΚΗΣ ΕΡΕΥΝΑΣ ΣΤΗΝ ΕΛΛΑΔΑ

Μελίνη Υφαντή-Κωστοπούλου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα

Σεπτέμβριος 2024





DEDICATION

Στην οικογένεια μου και τον Βασίλη





ACKNOWLEDGEMENTS

Πρώτα από όλα θα ήθελα να ευχαριστήσω την οικογένεια μου, τους γονείς μου και την αδερφή μου, τον παππού μου και την γιαγιά μου που μου έχουν σταθεί όλη μου την ζωή. Που πάντα πίστευαν σε εμένα και στις ιδέες μου, με παρακινούσαν να συνεχίζω και με έκαναν την γυναίκα που είμαι σήμερα.

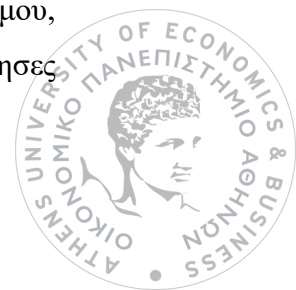
Επίσης θέλω να ευχαριστήσω τον καθηγητή μου και υπεύθυνο της διπλωματικής μου, κύριο Παπασταμούλη, που όταν του εξέφρασα την επιθυμία μου να κάνω μια διπλωματική σχετική με τα εκλογικά δεδομένα και την πολιτική έρευνα, βρήκε την λύση και μου πρότεινε αυτό το θέμα το οποίο με βοήθησε πολύ να συνδυάσω τα δύο μεταπτυχιακά προγράμματα στα οποία σπουδάζω και τις επιστήμες που με ενδιαφέρουν. Επίσης τον ευχαριστώ για την καθοδήγηση και την βοήθεια όλη την χρονιά κατά την συγγραφή της παρακάτω εργασίας.

Ακόμα ευχαριστώ για την καθοριστική του συμβολή, τον Επικ. Καθηγητή στο τμήμα Κοινωνιολογίας του Παντείου Φώτη Μηλιένο, ο οποίος είχε την ιδέα για την αξιοποίηση των δεδομένων από την Ευρωπαϊκή Κοινωνική Έρευνα. Αλλά και για τις συμβουλές του, ειδικά στην προετοιμασία και στην επεξεργασία των δεδομένων.

Έπειτα ευχαριστώ τις φίλες μου, την Άννα, την Κατερίνα, την Ιωάννα, την Ναυσικά, την Άσπα, την Ελευθερία, που μαζί τους ξεχνιόμουν, χαλάρωνα και μπορούσα μετά να συνεχίζω πιο δυνατή. Κορίτσια σας ευχαριστώ για όλα τα γέλια και για όλη την στήριξη. Ευχαριστώ και την φίλη μου Καλλιόπη γιατί τις περισσότερες ώρες που πέρασα γράφοντας την διπλωματική μου ήταν στο καφέ που εργάζεται, πάντα με την συντροφιά της και την στήριξή της.

Ένα ιδιαίτερο ευχαριστώ σε δύο νέες φίλες οι οποίες συνεχώς με παρακινούσαν. Στην Νάσια που γνώρισα στο μεταπτυχιακό στην Στατιστική και η βοήθεια της ήταν κομβική στην ολοκλήρωση της διπλωματικής. Ευχαριστώ που μου στάθηκες και με βοήθησες και τα περάσαμε όλα αυτά η μια δίπλα στην άλλη. Και την Αγγελίνα, που την γνώρισα μόλις φέτος στο μεταπτυχιακό στην Πολιτική Επιστήμη και στάθηκε δίπλα μου σαν φίλη χρόνων, ενδιαφέρθηκε και με βοήθησε και ας μην καταλάβαινε τίποτα από αυτά που έκανα.

Ακόμα, δεν γίνεται να μην ευχαριστήσω την Λάρα, μια υπέροχη φίλη που γνώρισα μόλις ένα χρόνο πριν και χωρίς αυτή δεν θα τα είχα καταφέρει. Λάρα σε ευχαριστώ για όλα τόσο πολύ. Για όλες τις ώρες που περάσαμε διαβάζοντας σπίτι μου, για όλες τις μακαρονάδες και τα γλυκά που μας μαγείρευες. Ευχαριστώ που με άφησες



να σου δείξω την πόλη μου και να σου μιλήσω για αυτή και εσύ την αγάπησες με όλη σου την ψυχή. Ευχαριστώ που άκουγες τις ατελείωτες αναλύσεις μου για τις εκλογές και την πολιτική σκηνή στην Ελλάδα. Που μαζί ακούγαμε ραδιόφωνο και εσύ χαιρόσουν και ας μην καταλάβαινες τι έλεγαν. Που όταν αγχωνόμουν με τα μαθήματα και το διάβασμα εσύ ήσουν εκεί να τα συζητήσουμε και να βρούμε την λύση. Είσαι μια φανταστική γυναίκα και μια *υπέροχη φίλη*, ευχαριστώ για όλα.

Τέλος, θέλω να ευχαριστήσω τον σύντροφό μου τον Βασίλη που είναι δίπλα μου από την πρώτη μέρα των σπουδών μου. Γνωριστήκαμε στην αίθουσα Δ12 την πρώτη βδομάδα των προπτυχιακών μας σπουδών στην Στατιστική και από τότε είσαι πάντα δίπλα μου, στις άπειρες ώρες διαβάσματος, στα άγχη και στις χαρές. Ήσουν δίπλα μου όλες τις φορές που σκέφτηκα να τα παρατήσω ή ότι δεν είμαι αρκετά καλή, για να μου δώσεις δύναμη να συνεχίσω. Ευχαριστώ για τα απίθανα φοιτητικά χρόνια που ζήσαμε μαζί στην ΑΣΟΕΕ, ευχαριστώ που πιστεύεις συνεχώς σε μένα και κάθε μέρα με κάνεις έναν καλύτερο άνθρωπο.





VITA

I studied Statistics at the Athens University of Economics and Business and during my studies, the study of politics, the study of political life and the history of the country were my main interests beyond statistics. I want to explore the ways in which the combination of social and political sciences with statistics can be productive. After discussions with my professors and people in my environment, I have concluded that political analysis, social research and political polling are areas that combine my interests. My choice to study in the MSc in Statistics at AUEB came as a logical consequence of the above concerns as well as my choice to study in the MA in Political Science and Sociology at National and Kapodistrian University of Athens. I hope to combine these two disciplines and become a useful scientist for society.





ABSTRACT

Melini Yfanti-Kostopoulou

ANALYZING INCOMPLETE VOTING DATA: THE CASE OF EUROPEAN SOCIAL SURVEY ROUND 10 IN GREECE

September 2024

This work aims to profile nonresponders concerning their voting behavior in the 2019 national election in Greece using data from Round 10 of the European Social Survey. Through various methodological approaches, we highlighted the complex nature of nonresponse bias, and we identified key characteristics that distinguish nonresponders from responders. Our analysis explored the Missing At Random (MAR) assumption and emphasized the importance of understanding nonresponders to reduce and deal with nonresponse bias effectively. We create a logistic regression model that identifies significant covariates that influence nonresponse, such as political engagement and interest, socio-economic factors, and demographic characteristics. Additionally, the clustering analysis shows that nonresponders form distinct subgroups with varying political attitudes and opinions. The clustering analysis revealed that nonresponders are a heterogeneous group, that reflects the broader diversity of the population making it difficult to mitigate the problem of nonresponse bias. This study contributes to the ongoing discourse on nonresponse in electoral and social surveys and offers valuable insights and methodological advancements to better understand and address this persistent challenge in survey research.





ΠΕΡΙΛΗΨΗ

Μελίνη Υφαντή-Κωστοπούλου

ΑΝΑΛΥΟΝΤΑΣ ΕΛΛΙΠΗ ΔΕΔΟΜΕΝΑ ΠΟΥ ΑΦΟΡΟΥΝ ΤΗΝ ΨΗΦΟ: Η ΠΕΡΙΠΤΩΣΗ ΤΟΥ ΚΥΚΛΟΥ 10 ΤΗΣ ΕΥΡΩΠΑΙΚΗΣ ΚΟΙΝΩΝΙΚΗΣ ΕΡΕΥΝΑΣ ΣΤΗΝ ΕΛΛΑΔΑ

Σεπτέμβρης 2024

Η παρούσα εργασία έχει ως στόχο να σκιαγραφήσει το προφίλ όσων δεν απαντούν στην ερώτηση σχετικά με την ψήφο τους στις εθνικές εκλογές του 2019 στην Ελλάδα, χρησιμοποιώντας δεδομένα από τον 10ο γύρο της Ευρωπαϊκής Κοινωνικής Έρευνας. Μέσω διαφόρων μεθοδολογικών προσεγγίσεων, αναδείξαμε την πολύπλοκη φύση της μεροληψίας της μη-απόκρισης και εντοπίσαμε βασικά χαρακτηριστικά που διακρίνουν αυτούς που δεν απαντάνε από αυτούς που απαντάνε. Η ανάλυσή μας διερεύνησε την υπόθεση τυχαίας μη-απόκρισης και υπογράμμισε τη σημασία της κατανόησης των χαρακτηριστικών όσων δεν απαντάνε για την αποτελεσματική μείωση και αντιμετώπιση της μεροληψίας μη-απόκρισης. Δημιουργήσαμε ένα μοντέλο λογιστικής παλινδρόμησης που εντοπίζει παραμέτρους που επηρεάζουν τη μη απάντηση, όπως η πολιτική ενασχόληση και το πολιτικό ενδιαφέρον, οι κοινωνικοοικονομικοί παράγοντες και τα δημογραφικά χαρακτηριστικά των συμμετεχόντων στην έρευνα. Επιπλέον, η ανάλυση συστάδων δείχνει ότι οι μη αποκρινόμενοι χωρίζονται σε διακριτές υποομάδες με διαφορετικές πολιτικές στάσεις και απόψεις. Η ανάλυση συστάδων αποκάλυψε ότι οι μη αποκρινόμενοι αποτελούν μια ετερογενή ομάδα, η οποία αντανάκλα την ευρύτερη ποικιλομορφία του πληθυσμού καθιστώντας δύσκολο τον μετριάσμο του προβλήματος της μεροληψίας της μη απάντησης. Η παρούσα μελέτη συμβάλλει στη συνεχιζόμενη συζήτηση σχετικά με τη μη απόκριση στις εκλογικές και κοινωνικές έρευνες και προσφέρει γνώσεις για την καλύτερη κατανόηση και αντιμετώπιση αυτής της επίμονης πρόκληση.



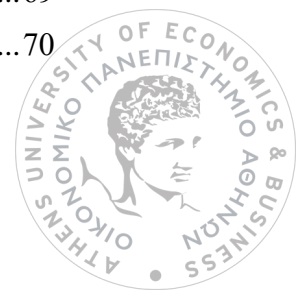


TABLE OF CONTENTS

CHAPTER 1	1
Introduction – Understanding Missing Data	1
1.1. Introduction: Setting the Stage.....	1
1.2. Non – Response Rates and Bias: The Looming impact.....	4
1.3. Exploring Types of Missingness	6
1.4. Identifying Missing Data Patterns	6
1.5. Unveiling Missing Data Mechanisms	7
1.6. Conclusion	8
CHAPTER 2	9
Literature Review	9
2.1. Introduction: The Nonresponder problem.....	9
2.2. Uncover Nonresponder Characteristics.....	9
2.3. Social Desirability Bias and the Spiral of Silence: Persistent Problems.....	10
2.4. Let’s Talk About Vote Overreporting	11
2.5. Predicting Vote Choice: Methods and Challenges	11
2.6. Conclusion.....	13
CHAPTER 3	15
Assumptions - Hypotheses	15
3.1. Assumptions - Hypotheses	15
CHAPTER 4	17
Data and the European Social Survey	17
4.1. Introduction	17
4.2. The European Social Survey: An Overview	17
4.3. Data Preparation.....	19
4.3.1. Continuous or Factors: Variable Treatment	22
4.3.2. Some Descriptive Statistics	22
4.4. Conclusion.....	28
CHAPTER 5	29
Methodology	29
5.1. Introduction	29
5.2. Strategies for Handling Missing Values	29



5.2.1 Handling Categorical variables	33
5.2.2. Computational algorithms (King et al., 2011)	34
5.2.3. Why choose multiple imputation?	35
5.3. Logistic regression and Variable selection	36
5.3.1. LASSO and GROUP LASSO.....	37
5.3.2. Stepwise variable selection via AIC	40
5.4. Logistic Regression Diagnostics	40
5.4.1. Receiver Operation Characteristic (ROC) Curve	41
5.4.2. K- fold cross validation	41
5.5. Clustering Techniques with Missing Data.....	41
5.6. Conclusion.....	43
CHAPTER 6	45
Implementation – Results	45
6.1. Introduction	45
6.2. Dealing with missing data	45
6.3. Variable selection using LASSO and Group – Lasso.....	46
6.4. Stepwise variable selection using AIC.	50
6.5. Evaluating Model Performance with ROC Curves	51
6.6. Profiling Non – Responders: Insights from the Optimal Model	53
6.7.1. Implementation	55
6.7.2. Interpreting Clustering Results	57
6.8. Conclusion.....	61
CHAPTER 7	63
Conclusion	63
7.1. Introduction	63
7.2. Overcoming Research Difficulties	63
7.3. An Overview of Data Preparation Process	64
7.4. Evaluation of the Methods	65
7.5. Was it Worth the Wait? Analyzing the Results and Validating the Assumptions.	67
7.5.1. Step 1: Logistic Model Results.....	67
7.5.2. An in-depth exploration of this special group	68
7.5.3. Realistic Scenario or Elusive Goal: Predicting the Vote of Nonresponders	69
7.6. Conclusion of the Conclusion	70



References	73
Appendix	79
List of tables.....	79
List of figures.....	86
An Overview of the Political Parties in Greece.....	91





LIST OF TABLES

Table 4.1: Recoding of the <i>edrec</i> variable.....	79
Table 4.2: Basic descriptive statistics for age.....	22
Table 4.3: Basic descriptive statistics for socio-political questions.....	24
Table 4.4: Descriptive statistics of the numeric variables, we treat scale variables ranging from 0 to 10 as numeric. We present mean, standard deviation, median, range and skewness.....	80
Table 6.1: The optimal lambda value λ_{\min} and the λ_{1se} for LASSO.....	47
Table 6.2: The optimal lambda value λ_{\min} and the λ_{1se} for Group-Lasso	48
Table 6.3: The optimal lambda value λ_{\min} and the λ_{1se} for LASSO.....	49
Table 6.4: The optimal lambda value λ_{\min} and the λ_{1se} for Group-Lasso.....	50
Table 6.5: Model 1 - assume ordinal as continuous – LASSO- stepwise.....	81
Table 6.6: Model 2 - ordinal as continuous - group lasso- stepwise.....	82
Table 6.7: Model 3 - ordered as factors- LASSO– stepwise.....	84
Table 6.8: Model 4 - ordered as factors- group lasso – stepwise.....	85
Table 6.9: Characteristics of participants that reduce nonresponse probability.....	54
Table 6.10: Characteristics of participants that increase nonresponse probability.....	54
Table 6.11: Demographic characteristics of nonresponders.....	54
Table 6.12: Estimated posterior distribution of the number of clusters.....	57
Table 6.13: Estimated number of observations per cluster conditionally on $K=2$	57
Table 6.14: Two - way frequency table "party voted for" vs "cluster".....	57
Table 6.15: Distribution of non - responders in the two clusters.....	58
Table 6.16: Characteristic of participants in cluster 1 according to the difference in the probabilities of choosing a level j from a covariate between the two clusters.....	59
Table 6.17: Characteristic of participants in cluster 2 according to the difference in the probabilities of choosing a level j from a covariate between the two clusters.....	60





LIST OF FIGURES

Figure 1.1: Party voted in last national election (whole sample).....	3
Figure 1.2: Party voted for in last national election (after listwise deletion).....	3
Figure 1.3: Party voted for in last national election 2019 (true results).....	4
Figure 1.4: Three Relevant Causal Models Linking Response Propensity with Nonresponse Bias (Groves & Peytcheva, 2008).....	7
Figure 4.1: Frequency bar plot showing the percentages of responders in each category of <i>edrec</i> variable.....	21
Figure 4.2: The distribution of the response variable Y.....	21
Figure 4.3: Histogram and Bar-plot for demographic variables 1/2.....	23
Figure 4.4: Histogram and Bar-plot for demographic variables 2/2.....	24
Figure 4.5: Bar-plots for variables that come from socio-political questions 1/3.....	26
Figure 4.6: Bar-plots for variables that come from socio-political questions 2/3.....	27
Figure 4.7: Bar-plots for variables that come from socio-political questions 3/3.....	28
Figure 4.8: Histograms of independent variables treated as numeric 1/4.....	86
Figure 4.9: Histograms of independent variables treated as numeric 2/4.....	86
Figure 4.10: Histograms of independent variables treated as numeric 3/4.....	87
Figure 4.11: Histograms of independent variables treated as numeric 4/4.....	87
Figure 4.12: Bar plots for independent variables treated as ordered factors 1/4.....	88
Figure 4.13: Bar plots for independent variables treated as ordered factors 2/4.....	88
Figure 4.14: Bar plots for independent variables treated as ordered factors 3/4.....	89
Figure 4.15: Bar plots for independent variables treated as ordered factors 4/4.....	89
Figure 4.16: Bar plots for independent variables treated as ordered factors.....	90
Figure 6.1: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for LASSO (assuming ordinal covariates as continuous).....	47
Figure 6.2: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for Group-Lasso (assuming ordinal covariates as continuous).....	48
Figure 6.3: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for LASSO (assuming ordinal covariates as factors).....	49
Figure 6.4: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for Group-Lasso (assuming ordinal covariates as factors).....	50
Figure 6.5: ROC curves for 5-fold cross validation (assuming ordinal covariates as factors). Model 1 after lasso and stepwise variable selection.....	52

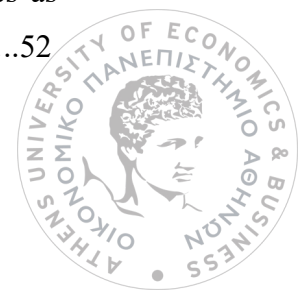


Figure 6.6: ROC curves for 5-fold cross validation (assuming ordinal covariates as factors). Model 2 after group – lasso and stepwise variable selection.....	52
Figure 6.7: ROC curves for 5-fold cross validation (assuming ordinal covariates as factors). Model 4 after group – lasso and stepwise variable selection.....	53
Figure 6.8: ROC curves for 5-fold cross validation (assuming ordinal covariates as factors). Model 4 after group – lasso and stepwise variable selection.....	53
Figure 6.9: The probability of choosing the j level of each categorical variable in each of the two clusters.	60



CHAPTER 1

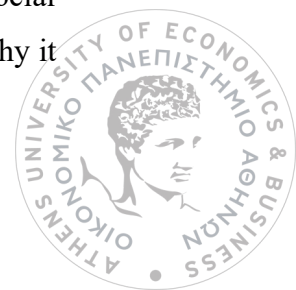
Introduction – Understanding Missing Data

1.1. Introduction: Setting the Stage

One of the primary challenges confronting social and political research is the issue of nonresponse. Despite the increased number of surveys in recent years, the response rate has been on a decline (Heer, 1999). Social scientists and statisticians delve into the reasons for this phenomenon, recognizing that the solution lies at the intersection of these two disciplines. Our focus centers on nonresponse related to questions about the voting behavior of interviewees. In this thesis, we aim to analyze and address this problem, employing both mathematical analysis and highlighting the importance of a sociological point of view.

In recent years, a global and national discourse has emerged, questioning the reliability of political and social surveys. This skepticism became more and more popular, particularly as opinion polls faced challenges in accurately predicting outcomes in crucial electoral events. In Greece, notable instances include the inaccurate estimations during the "double electoral earthquake of 2012" (Βούλγαρης & Νικολακόπουλος, 2014), the failure to foresee the results of the July 2015 referendum, the September 2015 elections, and the inability to accurately predict SYRIZA's share in the last parliamentary elections. An underlying factor contributing to these discrepancies may be the escalating nonresponse rates observed in recent years (Hoek & Gendal, 1997), with approximately half of survey participants choosing not to answer one or more questions (King et al., 2001).

Particularly in questions related to people's voting, a substantial percentage of undecided voters is observed in pre-election surveys or a high refusal rate in post-election surveys. The sample units refusing to answer voting questions are not negligible, and their removal from the sample can introduce serious bias, resulting in information loss (King et al., 2001) and, ultimately, a flawed estimate of the party share. To elucidate this, research indicates that individuals refusing to answer such questions share common characteristics. Nonresponse leads to the under-representation of this specific social group within the sample, introducing exclusion bias (Berinsky, 2002). In essence, we believe that nonresponse is not random but is influenced by social characteristics and political beliefs measured by other survey questions. This is why it



is crucial to identify those characteristics and create the profile of nonresponders to weight for this specific group in order to avoid bias.

It's important to understand that item nonresponse in political and social surveys is not random but rather a deliberate refusal. Deleting those observations would mean losing valuable information that they could provide in the sample. Additionally, research have shown that, while it's possible to entirely exclude this group from polling samples, reductions in the percentage of undecided participants may not necessarily result in improved accuracy of the poll estimates (Hoek & Gendal, 1997). In the same research Hoek and Gendal (1997) claim that the subsample with the biggest proportion of undecided responders produced the most accurate estimates. Their conclusion is that when allocating the undecided responders in proportion to party support levels together with direct intention questions, coupled with minimized contextual inquiries are probable to produce the most accurate poll estimates.

The issue of nonresponse to social and political surveys manifests itself in two ways: total refusal to participate in the survey, i.e., the absence of sample units (unit nonresponse), and participation but refusal to answer specific questions (item nonresponse). Both facets are equally grave and can introduce significant challenges to our estimates. Notably, there is a discernible gap in the literature concerning item nonresponse, particularly in the context of questions related to voting. Conversely, there exists a wealth of scientific articles addressing unit nonresponse in social surveys, particularly within the European Social Survey, the focus of our study.

To make clear our research question, we offer a graphical illustration. Research indicates that approximately 94% of researchers opt for listwise deletion, leading to the removal of entire observations. This approach, while famous, leads to a loss of crucial information and introduces bias into the analysis (King et al., 2001). Presented below is an estimation of the 2019 general election percentage based on responses from participants in the European Social Survey Round 10. A comparison between the survey estimates¹ (Figure 1.1) and the actual election results (Figure 1.3) reveals significant differences. It is essential to point out that when using the listwise deletion method (Figure 1.2), where responders who refuse to answer are removed from the sample, the accuracy of the estimate is not enhanced.

¹ The results in Figure 1.1 and Figure 1.2 are weighted by the variable: pspwght - post-stratification weight including design weight.



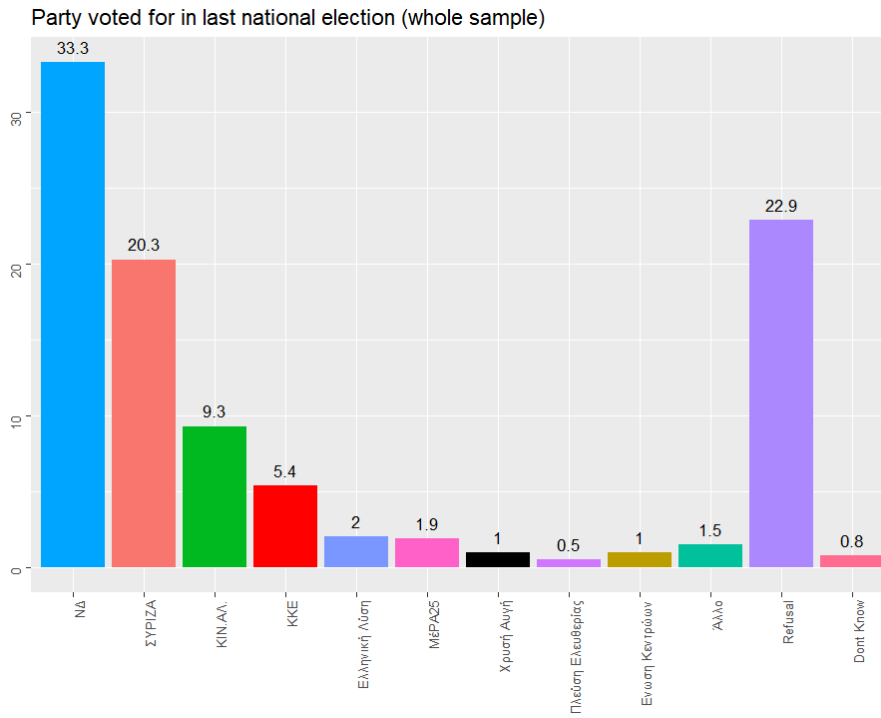


Figure 1.1: Party voted in last national election (whole sample)

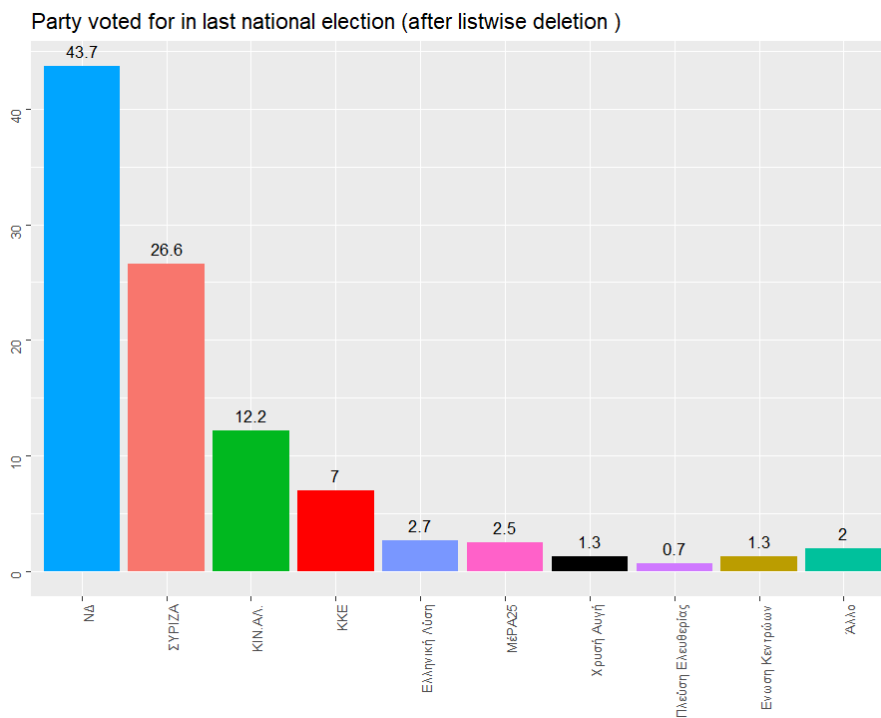


Figure 1.2: Party voted for in last national election (after listwise deletion)



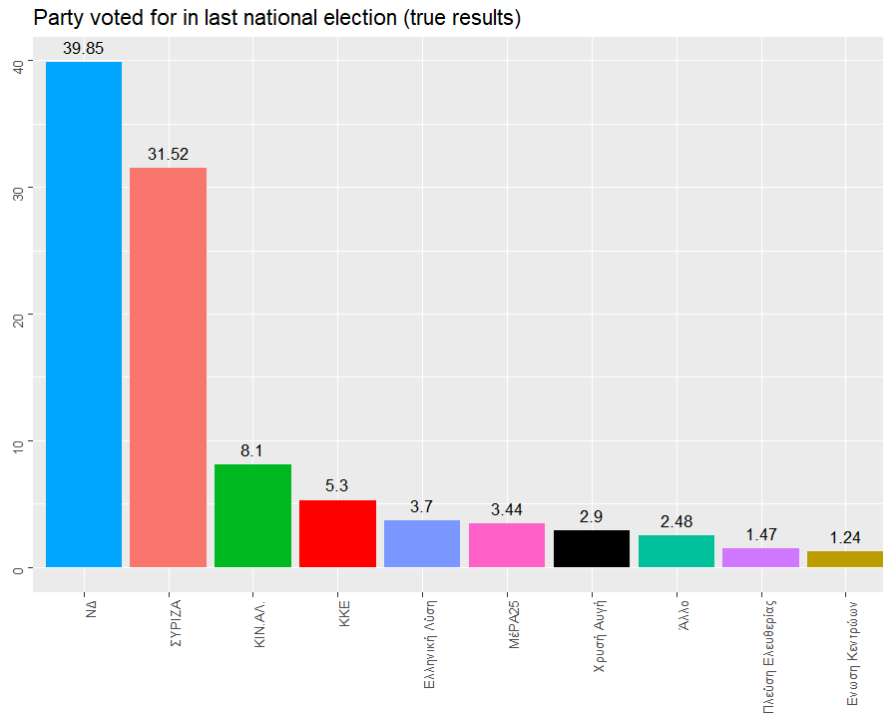


Figure 1.3: Party voted for in last national election 2019 (true results)

1.2. Non – Response Rates and Bias: The Looming impact

As Peytchev (2013) smartly explains : “Using the response rate as the only measure of the representativeness of a survey is erroneous because it is nonresponse bias that is feared, not nonresponse itself”.

First, we need to explain and describe nonresponse bias. It occurs when nonresponders have significantly different characteristics compared to responders. Since people who decide to withdraw their opinion may differ, from those who choose to participate, in their answers, this can lead to false estimation in population characteristics because of the underrepresentation of nonresponders characteristics in the sample. The above bias is common in survey studies, especially large-scale ones (Turk et al., 2019). We provide the definition of unit non – response bias for an estimate of the mean (Groves as cited in Peytchev, 2013):

$$Bias(\bar{y}_r) = E \left[\frac{m_s}{n_s} (\bar{y}_r - \bar{y}_m) \right] \quad (1)$$

According to commonly employed formulas, the nonresponse bias is the expected value of the difference between the response and the full sample means; in formula (1) the nonresponse bias is expressed as the product of nonresponse rate: total number of nonresponders (m_s) over the sample size (n_s), and the difference between responders

(\bar{y}_r) and nonresponders means (\bar{y}_m) (Peytchev, 2013). Formula (1) underlines the relation between response rate and nonresponse bias, the formula can lead us to the conclusion that increasing response rate will invariably decrease nonresponse bias (Peytchev, 2013).

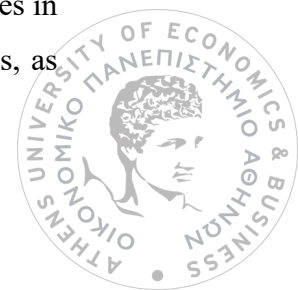
Another portrayal of nonresponse bias for a responder's mean could be the following, presented by Bethlehem (2002):

$$\sigma_{yp} / \bar{p} \quad (2)$$

- σ_{yp} is the covariance between survey variable y and the response propensity p .
- p is the response propensity, meaning the likelihood of participating in the survey conditional to the applied survey protocol.
- \bar{p} is the expected propensity over the sample members to be measured (response rate).

In formula (2), Peytchev (2013), explains that it is more clear that the covariance in the numerator is not independent of the response rate in the denominator and as a consequence an increase in the response rate may not necessarily mean a decrease in nonresponse bias and in the same time a lower response rate does not lead to an increased nonresponse bias.

In survey analysis when we use the sample to make inferences about a population, we assume that the sample which is analyzed is representative of the population, therefore both the sample and the responders need to be representative. It is common in social sciences to treat response rates as an indicator of data quality. Scientists tend to believe that higher response rates will lead to reduction of nonresponse bias and produce more accurate estimates. The essence of representative inference is that the characteristics of responders in a survey need to represent the characteristics of the target population (Fulton, 2018). It's true that a survey which has a high nonresponse rate is probable to be less representative of the population and suffer of nonresponse bias (Groves, 2006). It's crucial for our study to underline that nonresponse bias depends on the degree to which responders are different from nonresponders rather than being in direct relation to response rates. Moreover, the range of nonresponse bias depends on the response rate and the distinctiveness of nonresponders at the same time (Rogelberg & Stanton, 2007). Since nonresponse bias is not related to response rates in survey estimates across different studies, response rate is a poor indicator of bias, as



studied by Grooves and Petycheva (2008). Additionally, Peytchev (2013) claims that unit nonresponse could result to nonresponse bias. He underlines that as the response rate decreases the differences between responders and nonresponders lead to nonresponse bias without meaning that low response rates necessarily lead to significant nonresponse bias. If there are no differences between responders and nonresponders, then the sample is representative, and the inferences are valid (Anseel et al., 2010; Groves et al., 2006). Low response rates, influence survey estimates in the terms of bias, when responders and nonresponders differ in characteristics which are important for the topic we are studying (Fulton, 2018).

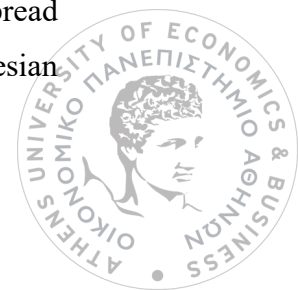
The above underlines the importance of understanding and creating nonresponders' profile. First and foremost, we need to prove that nonresponse is related to variables in our survey, this would mean that there are differences between the characteristics of responders and nonresponders. Additionally, knowledge of nonresponders' group special characteristics can lead to weight for them or to take additional measures in order to contain this type of people in our samples.

1.3. Exploring Types of Missingness

Two types of missingness are outlined in the bibliography: unit nonresponse and item nonresponse. Unit nonresponse occurs when someone refuses to participate in the survey and there is a total absence of this unit from the sample. Our focus is on item nonresponse, which, according to the literature, occurs when survey participants complete part of the survey but leave one or more questions unanswered (Graham, 2012). We are particularly interested in cases where nonresponse can be characterized as intentional refusal, as explained by Graham in *Missing Data: Analysis and Design* (2012): "It could be that the person leaves the question blank because of the fear that harm may come to him or her because of the response. It could be that the person leaves the questions blank because the topic is upsetting."

1.4. Identifying Missing Data Patterns

In literature we can find definitions about missing data patterns and missing data mechanisms. A missing data pattern shows how the observed and missing values are spread out in a dataset. It's different from a missing data mechanism, which explains why values might be missing. Essentially, patterns tell us where the missing data are, while mechanisms tell us why they're missing. In our dataset, missing values are spread across the entire data matrix. The three main methods—maximum likelihood, Bayesian



estimation, and multiple imputation—are all effective with this type of data configuration. Therefore, there's usually no need to select an analytical method only based on the missing data pattern (Enders, 2022).

1.5. Unveiling Missing Data Mechanisms

The missing data mechanisms or processes show us various ways in which the probability of the missing value is associated to the data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The MCAR mechanism means that the likelihood of missing values is independent of the data, while an MAR process suggests that missingness is associated only with the observed parts of the data. Conversely, an MNAR mechanism allows missingness to be influenced by the unobserved scores (Enders, 2020, p:6-11). The MCAR assumption is not very common. According to King et al. (2001): “if responders refuse to answer a vote preference or partisan identification question, then the data are not MCAR”. If the difference between responders and nonresponders to the vote preference can be predicted using variables in the dataset, the process would be MAR. Thus, the analyst can get closer to the MAR assumption by introducing more variables into the imputation process to predict the pattern of missingness. Lastly, in the case of MNAR (or nonignorable), other variables in the dataset cannot predict the missing values (King et al., 2001). It is crucial to note that listwise deletion may introduce bias unless MCAR holds (King et al., 2001). There is a relation between nonresponse bias and missing data mechanisms that could help us clearly understand the problem. Groves and Peytcheva (2008) graphically show in the figure we present below this relation.

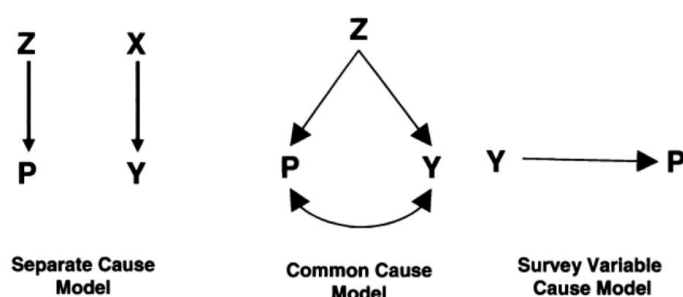


Figure 1.4: Three Relevant Causal Models Linking Response Propensity with Nonresponse Bias (Groves & Peytcheva, 2008).

As Groves (2006) explains, when turning from response propensities to nonresponse bias, a set of causal models is created. The MCAR corresponds to the first

model, the “separate cause” which assumes that the vector of causes of the Y variables is independent of the causes of response propensity P. This model would produce zero covariance in formula (2). This means that the expected values of Y among responders would be unbiased estimates of those among all sample participants. The “common cause” model corresponds to the MAR case, and shows that there are shared causes, Z of response propensity and the Y variables. The “common cause” model would produce zero covariance in formula (2) only when controlling for Z. Lastly the “survey variable cause corresponds to the MNAR case, and it asserts that Y itself is a cause of response propensity. The last model would produce a non-zero covariance (Groves & Peytcheva, 2008).

We will analyze item nonresponse patterns, assuming both refusals and "don't know" responses as refusals, in the question about the responders vote in the last national elections. Our initial assumption is that our data exhibit Missing at Random (MAR) characteristics, and our objective is to validate this assumption. To achieve this, we will employ a logistic regression model and conduct variable selection to pinpoint the variables correlated with the missingness. The dependent variable in our model, denoted as Y with values $\{0,1\}$, signifies 0 for observed observations in the “which party did you vote for” variable and 1 for missing observations.

We will validate that there are differences between responders to the vote question and non – responders and thus we are in danger of nonresponse bias. After creating the profile of nonresponders, it could be possible to control for Z and achieve decrease in nonresponse bias.

1.6. Conclusion

In conclusion, several questions from the ESS Round 10 questionnaire will serve as independent variables and we will try to find which of them are significant and influence the probability of someone being non-responder. Variable selection will be executed through a screening technique that zeroes non-important coefficients. Subsequently, we will construct the profile of those individuals who refused to answer what they voted for in the last national election (2019) and we will try to cluster them in categories.



CHAPTER 2

Literature Review

2.1. Introduction: The Nonresponder problem

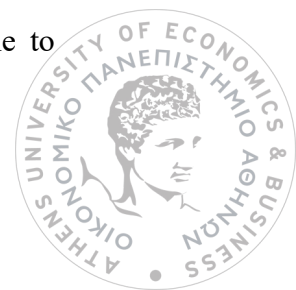
In this chapter we will discuss previous research results in the field of survey nonresponse. More specifically, we will present findings about nonresponders' characteristics and several approaches of classification of undecided voters and attempts of prediction of voting.

2.2. Uncover Nonresponder Characteristics

We will begin by formulating assumptions about the specific characteristics of nonresponders, drawing insights from existing literature on item nonresponse in social surveys. Numerous studies in the past have explored the characteristics associated with nonresponse, leading to the identification of certain factors. In a paper by John Brehm (1987), titled *Who's Missing? An Analysis of Nonresponse and Undercoverage in the 1986 National Election Studies Post-Election Survey* it is hypothesized that political interest and information interact with the likelihood of responding. Brehm (1987) does an analytical reference in previous research in this field. Specifically, income, race, and age have been linked to one's willingness to respond (O'Neil as cited in Brehm, 1987), and another study suggests that income, area of residence, and age are associated with the likelihood of not responding (DeMaio as cited in Brehm, 1987).

Furthermore, Alexander (2018) highlights that prior research consistently indicates that individuals from historically disadvantaged groups, including women, the impoverished, the elderly, and non-whites, are particularly prone to exhibiting item nonresponse behaviors such as responding with "don't know" or refusing to answer specific items (Adua & Sharp; Blom et al.; Candido et al.; Klein et al.; Kupek; Watkins & Melde; Wiederman ; Wiederman et al. as cited in Alexander 2018). These insights will serve as a foundation for our analysis of item nonresponse patterns related to voting behavior.

It is crucial to note that Alexander's (2018) examination of this hypothesis, particularly in the context of nonresponse to a question about political ideology, yields interesting findings. Contrary to the hypothesis that individuals from marginalized groups are less likely to answer specific questions, this study reveals that, apart from women and foreigners, people in such groups are not significantly more prone to



refusing to answer a question related to political ideology. Another research by Matsuo et al. (2018) also extends to different EU countries, where surveys are conducted to explore nonresponse to the European Social Survey (ESS). In some cases, a supplementary questionnaire is administered to those who refused to answer. For instance, research in Belgium and Norway shows that cooperative responders in Belgium are less likely to be in the age group of 30-39 or 60 and above compared to the 14-29 age group. Cooperative responders in Belgium also participate more in social activities and exhibit a higher interest in politics. In Norway, cooperative responders are more likely to be highly educated, engage in social activities, express satisfaction with the functioning of democracy, and hold positive views about immigrants (Matsuo et al., 2018). These insights contribute to our understanding of nonresponse patterns in the context of the European Social Survey.

2.3. Social Desirability Bias and the Spiral of Silence: Persistent Problems

The phenomenon of responders answering "I don't know" or misreporting their vote, potentially influenced by what they believe the interviewer wants to hear is strongly associated with the concept of social desirability bias (Valentino et al., 2017). Social desirability bias suggests that responders, especially those with more "extreme" opinions, may alter their answers to conform to socially acceptable norms (Edwards, 1953). Moreover, the interviewer effect introduces an additional layer of social desirability bias, as individuals may shape their responses based on their perception of the interviewer's expectations or observable traits, particularly when the survey topic is associated with characteristics like social class, gender, or ethnic background (Krumpal, 2013).

Another theory which can help explain item nonresponse in the context of social and political surveys, particularly when respondents choose not to disclose their political and vote preferences, is the spiral of silence theory. This theory, proposed by Noelle-Neumann in 1974, posits that individuals are less likely to express their opinions publicly if they believe that their views are in the minority. This can lead to underrepresentation of minority opinions. We understand that if individuals sense that their political stance is unpopular or controversial, they may opt for nonresponse to avoid potential social backlash or judgment. This dynamic can affect the accuracy and completeness of survey data, especially on sensitive political topics like voting behavior.



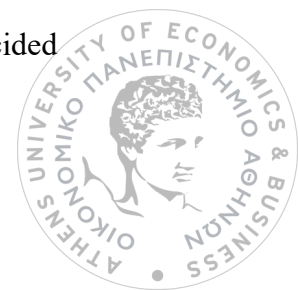
2.4. Let's Talk About Vote Overreporting

Another intriguing aspect is the phenomenon of vote overreporting in postelection surveys, where a higher proportion of the population claims to have voted compared to official vote records (Silver et al., 1986). Again, this suggests that responders choose the more socially acceptable answer. In the context of our research, where participants may overreport voting, exploring whether these responders share similar characteristics with those who refuse to answer the question about their voted party in the last national elections or prefer to answer "I don't know" could provide valuable insights. It raises the question of whether certain responder traits or attitudes contribute to both overreporting and nonresponse in subsequent political questions.

2.5. Predicting Vote Choice: Methods and Challenges

The question arises: what happens once we've mapped out the characteristics of item nonresponders? Knowing the characteristics of nonresponders in vote related questions serves several purposes. Firstly, it allows us to apply methods aimed at correcting for nonresponse bias and adjusting weights based on the particular characteristics of the group of individuals who opt not to respond. Additionally, we can proceed to classify nonresponders, identifying distinct groups, and attempt to estimate their past or future voting behavior. Previous research has addressed both aspects, and we will provide a brief commentary on them. It's worth noting that most of the scientific literature focuses on pre-election surveys, given our interest in predicting responders' voting intentions. However, these methods can also be extended to post-election surveys.

In literature a two stages method for classifying undecided voters is proposed by Fenwick et al. (1982). In order to predict the potential voting behavior of each undecided voter, the authors claim that a two-step process is necessary. In Step 1 a model that elucidates the voting preferences of decided voters, utilizing attitudinal, candidate evaluation, and demographic variables, is estimated. A decided voter is defined as someone who, during the survey, expressed an intention to vote for a specific candidate. In Step 2 we use this model to formulate a decision-making criterion for classifying undecided voters. Essentially, this criterion examines the attitudinal, candidate evaluation, and demographic data of each undecided voter and compares it to the typical response pattern of each candidate's decided voters. Undecided voters are then projected to be most inclined to vote similarly to their most analogous decided



counterparts. Given that voter preference is a categorical or nominally scaled variable, multiple discriminant analysis emerges as a fitting analytical tool. This approach employs a collection of predictors (referred to as discriminating variables) to classify individuals into one of two or more mutually exclusive groups (Fenwick et al., 1982).

In literature we can also find the opinion that if we allocate randomly nonresponders to the possible choices, we can improve the predicting accuracy of our model. In their paper Visser et al. (2000) claim that if we assume undecided people vote randomly, it can make predictions even better. In fact, they say that adding 2 percent random responses to each candidate's expected votes in mail surveys made the predictions more accurate by reducing the expected winning margin.

In their study *A Bayesian Allocation of Undecided Voters*, Nandram and Choi (2008) analyze election poll data in the United States, particularly focusing on polls leading up to the November election. The data is typically presented in tables showing candidate preferences and voter statuses. For instance, in the 1998 Buckeye State Poll for the governor's race, conducted in January, April, and October, the tables categorize voters as likely or unlikely to vote for the candidates. However, a significant number of voters remain undecided across all polls. To address this, the authors use Bayesian methods to allocate these undecided voters to the various candidates. Their approach allows for modeling different scenarios of missing data, assuming both ignorable and non-ignorable patterns. They employ a multinomial-Dirichlet model to estimate probabilities within the tables, aiding in predicting the eventual winner. The study introduces a time-dependent model for non-ignorable nonresponses across the three polling periods. This model builds upon an ignorable nonresponse model, incorporating flexibility and uncertainty regarding the assumption of ignorability. Additionally, the authors compare their proposed model with two others: an ignorable and a non-ignorable nonresponse model that assumes a common stochastic process to leverage data across time. The authors utilize Markov chain Monte Carlo methods to fit these models (Nandram & Choi, 2008).

In their study titled *Modeling Undecided Voters to Forecast Elections: From Bandwagon Behavior and the Spiral of Silence Perspective*, Liu et al. (2021) explore the challenges in forecasting election results, particularly focusing on the neglected aspect of undecided voters in public opinion polls. Despite various forecasting methods, opinion polls remain prevalent, yet they often overlook the significant influence of



undecided voters. Acknowledging the impact of undecided voters on election outcomes, the authors analyze their potential behavior through the lens of the bandwagon effect and the spiral of silence theory. To address this, the authors develop a hierarchical Bayesian forecasting model aimed at predicting voting results. They apply this model to two significant events: the 2016 United States presidential election and the 2016 Brexit referendum. Their findings suggest that considering the impact of undecided voters enhances the predictability of voting outcomes. By integrating aggregated polls into the hierarchical Bayesian framework, the model proves to be a robust predictor of election results. Furthermore, their results underscore the importance of sentiment based on voter expectation in forecasting election outcomes.

2.6. Conclusion

Concluding this chapter, the literature is rich in the field of pre- election surveys and exit polls. Several methods contained supplementary questionnaires and follow up questions for nonresponders. Moreover, there is a lack of research in the field of nonresponse in vote related questions in post-election surveys. Allocation of undecided voters is a topic we can find in previous research but not in the range we were expecting.





CHAPTER 3

Assumptions - Hypotheses

3.1. Assumptions - Hypotheses

The literature review, led us to interesting hypotheses:

1. Nonresponse is not random: According to the literature nonresponse is influenced by sociopolitical factors. To be more precise, prior studies (Brehm, 1987; Alexander, 2018; Matsuo et al., 2018) have shown that nonresponse is often linked to specific demographic and socio-economic characteristics such as age, income, education and political interest.
2. We assume that socio-economic and political factors significantly influence nonresponse bias. Higher nonresponse rates among specific demographic groups, such as the economically disadvantaged or politically disillusioned, are expected to introduce bias in survey estimates if not properly addressed.
3. Political context and party dynamics affect response rates: The political climate in Greece, which is characterized by significant events like the “double electoral earthquake of 2012” and the referendum (Βούλγαρης & Νικολακόπουλος, 2014), is unique and complex. We assume that the political context and dynamics of individual parties significantly affect response rates. Voters' alignment with major parties and their stance on key issues likely influence their willingness to respond to questions.
4. We assume that our data belongs to the MAR case. We hypothesize that nonresponse is related to observed variables, allowing us to use these variables to predict for missing data. In essence, we expect to see distinct demographic profiles for responders and nonresponders, based on the logistic regression model. In addition, we expect the logistic regression model to perform better than a random prediction of non- response.
5. Given the different opinions expressed for nonresponders' profile in the literature, and the difficulty in constructing a solid profile, we assume that nonresponders are not a homogeneous group but can be categorized into distinct clusters based on their responses to other survey questions. We assume that these clusters will provide deeper insights into the potential voting behavior of nonresponders, helping to mitigate the bias introduced by missing data.





CHAPTER 4

Data and the European Social Survey

4.1. Introduction

In the chapter below we will explain the source of our data. We will provide information about the European Social Survey, their objectives, and the way they gather the data. At the same time, we will present how we “cleared” our data set by deleting variables that weren’t useful for our analysis and at the same time creating new variables and combining old ones. We will also mention our approach of handling ordinal variables as numeric or as factors.

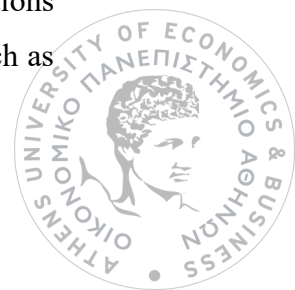
4.2. The European Social Survey: An Overview

The data for our analysis is sourced from the European Social Survey (ESS), specifically from Round 10 initiated in 2020. The ESS is a comprehensive pan-European, academically driven social survey that has been conducted across numerous European countries since 2001. Greece has actively participated in five rounds of the survey. The ESS aims to capture attitudes, perceptions, and behaviors across diverse populations in more than thirty countries, providing valuable insights into societal dynamics (European Social Survey, n.d.).

The primary objectives of the ESS are as follows:

1. To systematically observe and interpret the dynamics of social structure, conditions, and attitudes in Europe, analyzing how the continent's social, political, and moral landscape evolves over time.
2. To establish and promote elevated standards of precision in cross-national research within the social sciences, encompassing elements such as questionnaire design and pre-testing, sampling, data collection, bias reduction, and question reliability.
3. To introduce reliable indicators of national progress derived from citizens' perceptions and evaluations of crucial aspects of their societies.
4. To undertake and facilitate the training of European social researchers in the field of comparative quantitative measurement and analysis.
5. To enhance the visibility and dissemination of data on social change, reaching academics, policymakers, and the broader public.

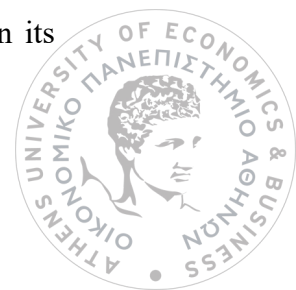
The questionnaire employed in the survey comprises a core set of questions repeated in each round, complemented by alternating sections covering topics such as



current affairs, human values, and test questions. Round 10 introduced two new sections focusing on Digital Social Contacts in Work and Family Life, and Understandings and Evaluations of Democracy. The questions in Round 10 are categorized into various sections, including:

1. Media and social trust
2. Politics
3. Subjective well-being, social exclusion, religion, national and ethnic identity
4. Understanding of democracy
5. Gender, Year of birth, and Household grid
6. Socio-demographics
7. Digital social contacts in work and family life
8. Human values
9. Test questions
10. Impact of COVID-19

In Greece, the National Centre for Social Research has been responsible for conducting the research. Greece has participated in 5 rounds of the ESS up to now (ESS1, ESS2, ESS4, ESS5, ESS10). The questionnaire is typically administered through face-to-face interviews, although, for the 10th round, countries were given the option to choose between responder-completed questionnaires and video interviews. The data collection process in Greece spanned from November 2021 to May 2022, reflecting the challenges and adaptations necessitated by the ongoing COVID-19 pandemic. The European Social Survey (ESS) adopts a comprehensive sampling strategy to ensure representation and inclusivity. The survey targets individuals aged 15 and above living in households in Greece, irrespective of nationality, citizenship, or language. The sampling approach is characterized by 3D stratified random sampling; a method designed to achieve an “optimal sample size” of at least 1500 individuals. This process considers various factors influencing the final sample design (European Social Survey, n.d.). The sampling frame is constructed based on the 2001 Population Census data, covering the entire population of Greece. The three-stage sampling process involves the random selection of surface area units in the first stage, private households in the second stage, and individuals within the selected households in the third stage. In the tenth round of the ESS in Greece, the survey achieved a response rate of 48.0%. It is worth noting that one of the primary objectives of the ESS, as outlined on its



website (European Social Survey, n.d.), is to establish and promote high standards for international research in the social sciences. This includes addressing issues related to questionnaire design, interviewer training, composite sampling, and weighting techniques. However, despite these efforts, there are notable challenges associated with nonresponse, particularly in the form of item nonresponse, which will be the focus of our analysis.

4.3. Data Preparation

We initiated our analysis by preparing our data, underscoring that the original ESS dataset for Greece encompassed 586 variables and 2799 observations. To streamline our approach to variable selection and model building, we needed to constrain the dimensions of our data. Initial steps involved removing auxiliary variables, interview mode-related variables, weights, and data collection specifics. Subsequently, control questions (test questions) were excluded, after we ensured consistency in each responder's answers. We also deleted question variables which were not posed to all participants or questions that were related to response in previous ones. Columns deemed irrelevant to our research question, such as those related to religion (we kept only *rlgdgr: How religious are you?*) and those addressing victims of discrimination (we kept only *dscrgrp: Would you describe yourself as being a member of a group that discriminated against in this country?*), were also eliminated. From the demographic and socio-demographic related questions, only basic information concerning the responder was retained. We deleted all variables related to socio-demographic characteristics of responder's household members and family.

We deleted the variable *health: How is your health in general? Would you say it is...* and the variable *anctry1: How would you describe your ancestry?* since we believe that they are irrelevant with the topic we are analyzing.

We omitted the variable *region: Region (country)* and we kept the variable *domicil: Which phrase on this card best describes the area where you live?* which describes the characteristics of responder's residence and contains the information we are interested in (if they live in a city or in a village etc.).

We discard the variable *lnghom1: What language or languages do you speak most often at home?* We performed Fisher's exact test in order to identify if there is a significant relation between the language a responder speak and the response variable



Y. Since the relation is not significant (p -value =0.466), we decided to discard this variable.

Concerning the income the data set has three questions.

- *hincsrca*: Please consider the income of all household members and any income which may be received by the household as a whole. What is the main source of income in your household?
- *hinctnta*: Household's total net income, all sources.
- *hincfel*: Feeling about household's income nowadays.

We decided to discard the first one. We believe that the two variables that we keep contain all the information we need for our analysis.

Simultaneously, we introduced new variables, including:

- *scltrst*: measuring social trust and derived as the sum of three questions related to social trust (range: (0-30)).
- *poleng*: measuring involvement in politics, computed as the sum of seven variables measuring political activity in the last 12 months (range: (0,7)).
- *hhmmb1*: resulting from recoding *hhmmb* (number of people living in the household), creating a dichotomous variable with the categories "one-person household" and "multi-person household" (Matsuo et al., 2018).
- *occptn*: obtained by recoding *mnactic* (main activity last 7 days), generating a dichotomous variable with the categories "employed" and "unemployed" (Matsuo et al., 2018).
- *edrec*: emerging from the recoding of *edlvegr* (highest level of education in Greece), introducing a new variable with six categories (out of 19). Education in the ESS adopts the International Standard Classification of Education (ISCED) categories, and we recoded the education level in Greece into four categories: lower secondary and less; upper secondary; advanced vocational; and tertiary (Piekut, 2019) (Table 4.1, Appendix). There is a small percentage of responders who belong to category other, to be more exact, in Figure 4.1 below we can see that only 0.3% of the responders answered 'other' in the education level question.



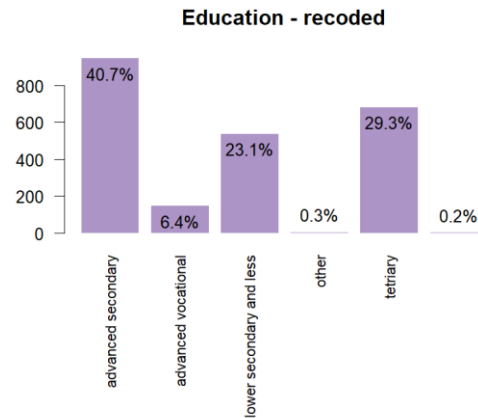


Figure 4.1: Frequency bar plot showing the percentages of responders in each category of *edrec* variable.

We decided to delete those responders since we think that the variability of the education status is covered from the answers provided. Additionally, we created the response variable for our study, represented by a dichotomous variable Y (0: answer, 1: not-answer), indicating whether the observation provided an answer to the question "*Which party did you vote for in the last national elections?*". In Figure 4.2 below we can see the distribution of the response Y.

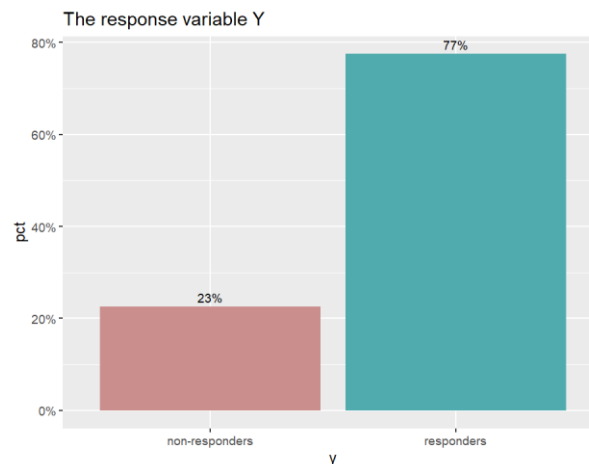


Figure 4.2: The distribution of the response variable Y.

As previously mentioned, we plan to execute logistic regression to identify which are the variables that significantly influence the nonresponse. To this end, responders who did not vote in the last national elections were excluded from our analysis.

4.3.1. Continuous or Factors: Variable Treatment

It is important to note that our covariates are nominal, ordered factors and integers. We decide to treat ordinal variables as both continuous and factors and identify the best strategy. We can treat ordinal variables as continuous if we can assume that successive categories of the ordinal independent variables are equally spaced (Long & Freese, 2006). Other researchers, suggest that treating ordered as continuous doesn't depend on the number of categories nor the marginal distribution of ordinal variables (Robitzsch, 2020). We will try both approaches and examine which is optimal.

4.3.2. Some Descriptive Statistics

To be able to understand better the characteristics of the sample we present some descriptive statistics. Since the final dataset contained 126 variables, we decided to discuss demographic variables and some socio-political opinions. The rest of the variables' descriptives can be found in the Appendix (Table 4.4 and Figures 4.8 – 4.16).

Starting with the demographic characteristics of the sample. In Figures 4.3 and 4.4 below we can see the illustration of the distribution of these variables. We make some comments:

- *agea*: After examining the histogram and creating Table 4.2, we observed that age ranges from 18 to 89. We explained above that the target population for ESS is from 15 years old. Since we discarded all participants that did not vote in national elections of 2019 it is normal for this variable to have as minimum the 18 years.

Basic descriptive statistics					
<i>Variable</i>	<i>Mean</i>	<i>SD</i>	<i>Median</i>	<i>Range</i>	<i>Skewness</i>
agea	51.61	15.90	52	71 (18-89)	0.08

Table 4.2: Basic descriptive statistics for age.

- *hinctnta*: Is the variable describing the household's income. We noticed that there is a large percentage (>40%) of nonresponse. We expected this to happen, in literature there is a rich discussion about nonresponse in income related questions.
- *ctzcntr*: This variable describes whether a participant is citizen of the country or not. The majority of the participant are citizens of Greece (category 1). There is a small percentage of the participants who are not citizens of Greece (category 2) and an even smaller percentage who refuse to answer this question.



- *gndr*: The gender distribution shows that the majority in the sample (in correspondence with the population) are women (category 2).
- *domicil*: This variable describes the area where the respondent lives. The majority of participants in the survey live in a big city (category 1), followed by country village (category 4), town or small city (category 3) and lastly suburbs or outskirts of a big city (category 2).
- *occpn*: This is a variable created by us. The majority of the participants in the survey are employed (category 1). The unemployed (category 2) percentage is not small, in Figure 4.4, we observe that it is over 40%.
- *hhmmb1*: This variable is also new, created by us and it shows the number of people living in a household. The majority of the participants live in a multi-person household.
- *lrscale*: We decided to include the variable that shows how participants place themselves in the scale where 0:left and 10:right. We observe that nonresponse is $\approx 10\%$ and the majority of participants place themselves in the center.

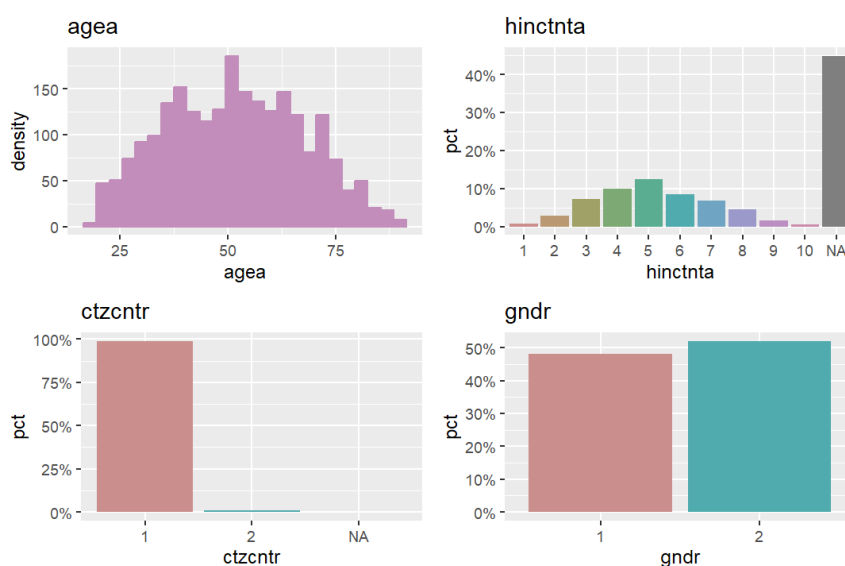


Figure 4.3: Histogram and Bar-plot for demographic variables 1/2

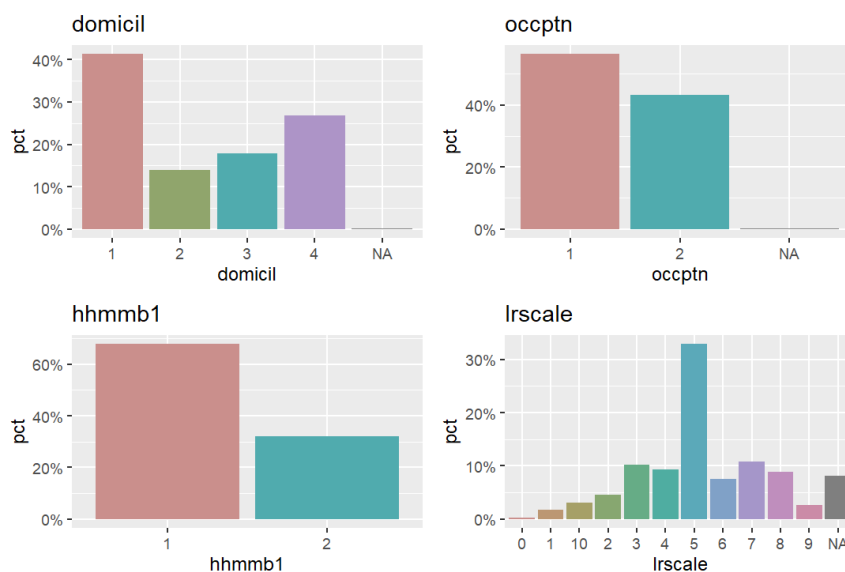


Figure 4.4: Histogram and Bar-plot for demographic variables 2/2

We continue to discuss about descriptive statistics, but now in terms of sociological and political questions. We selected 15 questions to investigate our sample's beliefs and opinions around several topics. We present graphically these 15 variables in Figures 4.5-4.7, and we make some comments:

- *stfgov*: This variable measures how satisfied is someone from the national government in a scale from 0:extremely dissatisfied to 10:extremely satisfied. As we observe in Figure 4.5 and in Table 4.3 the distribution is nearly symmetrical. The majority of the participants answer level 3 or 5. Moreover we observe that the “extremely dissatisfied” are more than the “extremely satisfied”.

Basic descriptive statistics					
<i>Variable</i>	<i>Mean</i>	<i>SD</i>	<i>Median</i>	<i>Range</i>	<i>Skewness</i>
stfgov	4.20	2.30	4	10 (0-10)	0.06
votedir	7.98	1.76	8	10 (0-10)	-0.98
fairelcc	7.32	2.11	8	10 (0-10)	-1.03
dfprtalc	5.72	2.44	6	10 (0-10)	-0.38
rghmgprc	5.91	2.35	6	10 (0-10)	-0.27
gptpelcc	5.65	2.71	6	10 (0-10)	-0.30
gvctzpvc	3.81	2.32	4	10 (0-10)	0.23
wpestopc	4.54	2.40	4	10 (0-10)	0.15
panfolru	5.07	2.24	5	10 (0-10)	-0.10

Table 4.3: Basic descriptive statistics for socio-political questions.

- *votedir*: This variable measures how important, a participant believes, that is for the democracy, citizens to have the final say on political issues by voting directly in referendums. Again, a scale, from 0: not at all important for democracy in general to 10: extremely important for democracy in general, is used. Most of the answers are above level 7. The distribution is left skewed.
- *fairelcc*: In this question participants express how much the statement that the elections are free and fair applies to their country. The scale that is used is from 0: does not apply at all to 10:applies completely. Figure 4.5 and Table 4.3 show that the distribution is left skewed and most of the observations are > level 7.
- *dfprtalc*: Participants needed to think how much the following statement applies to their country: “In country different political parties offer clear alternatives to one another”. Figure 4.5 and Table 4.3 reveal that more people believe that this statement does not apply at all compared to those who believe that it applies completely. Nevertheless, categories 6-9 gather more answers than categories 1-4.
- *rghmgprc*: The question measures how much participants think that this statement applies to their country: “In country the rights of minority groups are protected”. The distribution is almost symmetric. Those who believe that this applies completely are more than those who believe that it doesn’t apply at all.
- *gptpelcc*: The question measures how much participants think that this statement applies to their country: “In country governing parties are punished in elections when they have done a bad job”. The distribution is almost symmetrical.



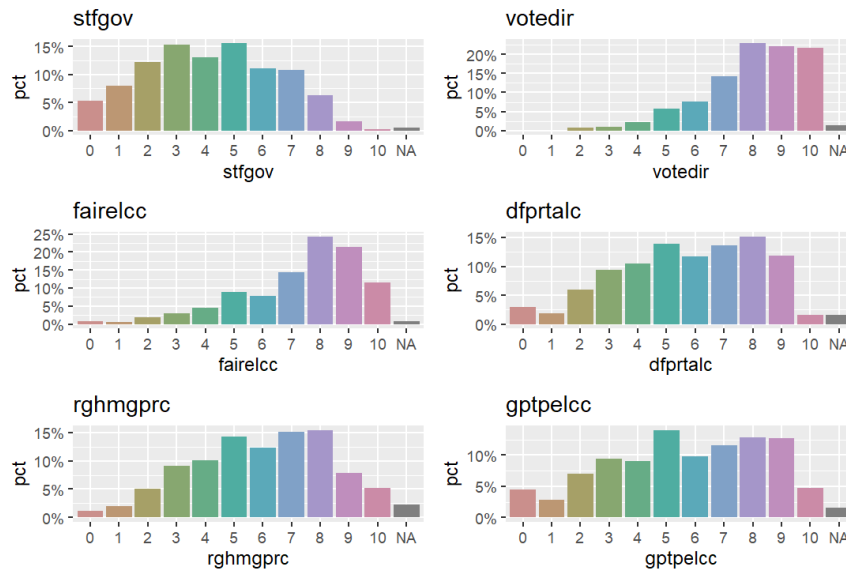


Figure 4.5: Bar-plots for variables that come from socio-political questions 1/3

- *gvctzpv*: The question measures how much participants think that this statement applies to their country: “In country the government protects all citizens against poverty”. After examining Figure 4.6 and Table 4.3 we can tell that the majority of participants seems to believe that this statement does not apply in Greece.
- *wpestop*: More participants tend to believe that in Greece the will of people can be stopped compared to those who believe that it can't.
- *accalaw*: When the participants were asked how acceptable it would be for Greece to have a strong leader who is above the law, the majority answered that this is not at all acceptable.
- *panforlu*: This variable measures if it is more important to follow government rules or to make own decisions when fighting a pandemic. They use a scale from 0: Much more important to follow government rules to 10: Much more important to make own decisions. The distribution is symmetrical.
- *freehms*: When the participants were asked if they agree that gays and lesbians are free to live life as they wish, the majority answered that they Agree Strongly or they Agree (levels 1 and 2 respectively).
- *loylead*: When the participants were asked if they agree that country needs most loyalty towards its leaders, the majority answered that they Disagree or Disagree strongly (levels 4 and 5 respectively)



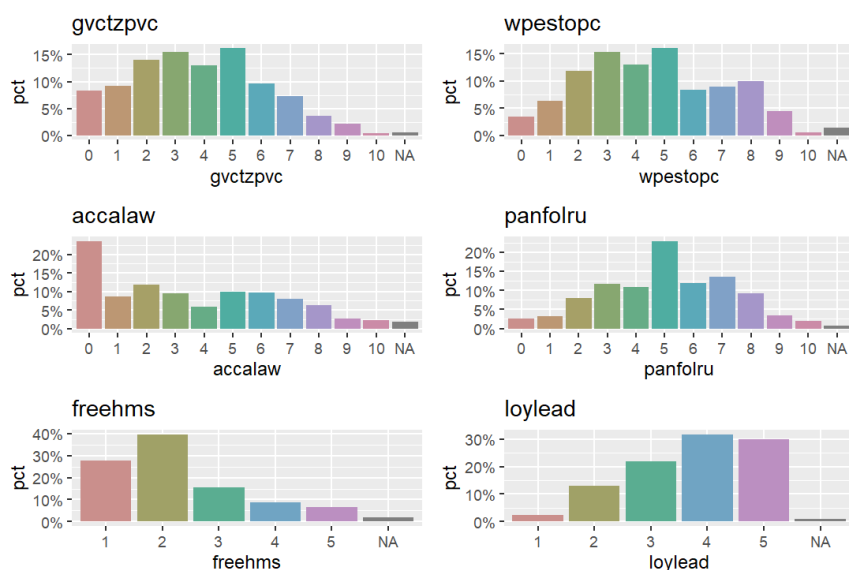


Figure 4.6: Bar-plots for variables that come from socio-political questions 2/3

- *ccnthum*: Participants were asked if they believe that climate change was caused by natural processes, human activity, or both. Most of the sample believed that it was caused about equally by natural processes and human activity (level 3) and that it was caused mainly by human activity (level 4).
- *gvconcl9*: The participants were asked if they agree that coronavirus is the result of deliberate and concealed efforts of some government or organization. The majority of the sample Disagree and Disagree Strongly.
- *dscrgrp*: The participants were asked if they would describe themselves as being a member of a group that is discriminated against in this country. The majority of participants asked that they are not member of a discriminated group.

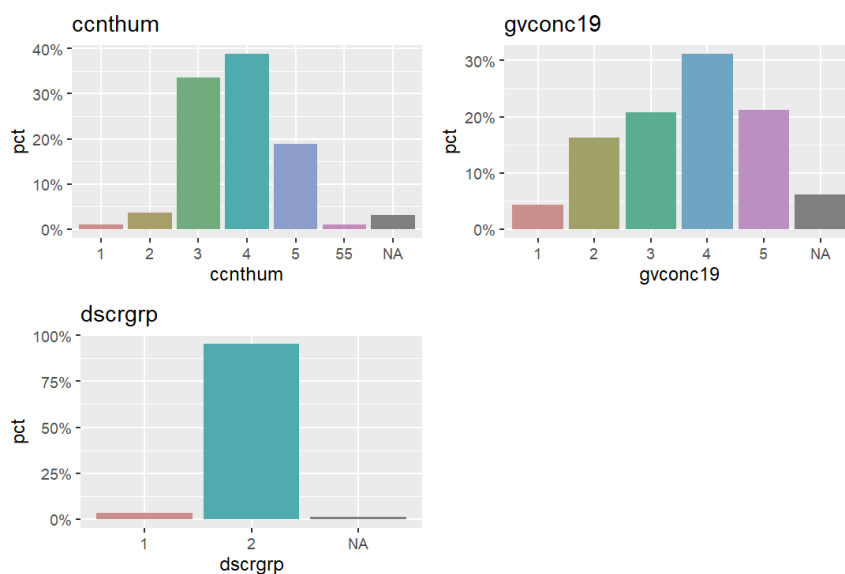


Figure 4.7: Bar-plots for variables that come from socio-political questions 3/3

Above we presented some descriptive statistics for a part of our variables. All the variables we presented contain missing values. The descriptive statistics for the rest of the variables can be found in the Appendix. In addition, the website of the ESS² provides description and graphs for all the variables and all the countries.

4.4. Conclusion

Concluding this chapter, our data are from the [European Social Survey Round 10](#). We excluded auxiliary variables, as well as variables that we decided were not informative for our analysis. We created new variables by combing or recoding already existing ones. We also created our binary response variable Y. Lastly, we discussed the types of our variables and how we should treat ordinal variables.

² <https://ess.sikt.no/en/datafile/f37d014a-6958-42d4-b03b-17c29e481d3d/262?tab=1>



CHAPTER 5

Methodology

5.1. Introduction

The methodology chapter aims to explain the path of our analysis. We elaborate each method's theoretical background and our decision to use it. We will start with multiple imputation in order to manage missing values in the independent variables, then we will proceed to logistic regression and variable selection using screening techniques like LASSO and Group – Lasso and stepwise variable selection and lastly, we will explain how we evaluated the model and which method we used to cluster nonresponders.

5.2. Strategies for Handling Missing Values

In the upcoming section, we will delve into the multiple imputation technique for managing missing data. Additionally, we will thoroughly explain our decision to utilize the [AMELIA](#) library in R (King et al., 2011) to implement multiple imputation and discuss our approach on how to handle nominal and ordinal variables within the imputation process.

It's crucial to note that our dataset harbored numerous missing values (including refusals and "don't know" responses) across almost all explanatory variables. Dealing with missing values in numeric, ordered, and nominal variables is intricate. Listwise deletion, as previously mentioned, poses a risk of significant information loss, given its deletion of both nonresponses and responses. Deleting the missing values would mean deleting 2/3 of our observations. We will assume that the data support the Missing at Random (MAR) pattern, this means that they can be predicted using other variables in the dataset. In political science surveys, missing data are generally unlikely to adhere strictly to the Missing Completely at Random (MCAR) assumption, as they invariably relate to political and sociological phenomena (Lall, 2016). Lastly the MNAR assumption will not add useful information since most data sets contain sufficient information (King et al., 2001).

We will employ multiple imputation to predict missing values, specifically utilizing the EMis method for multiple imputation (King et al., 2001). The EMis, multiple imputation approach becomes more and more popular among political



scientists due to its enhanced efficiency compared to listwise deletion and its potential for unbiased outcomes under correct distributions of missing data (Lall, 2016).

There are several methods to deal with missing values, one of them is listwise deletion which we already mentioned, another one is overall mean imputation. These two techniques can lead to inefficient analyses and produce biased estimates. This is why we choose to implement a more sophisticated technique, multiple imputation.

But what is imputation? When using imputation techniques, we predict missing units of a variable using the subject's other, known characteristics (Donders et al., 2006). As already mentioned, observations with missing data related to other known characteristics belong to the MAR case and by definition they are a random subset from the sample given these other known characteristics (Donders et al., 2006). As explained by Donders et al. (2006), these missing values could be replaced by randomly chosen values from the part of the sample that we can identify by these characteristics.

To define multiple imputation, we must say that we impute m values for each missing item and as a consequence we create m completed data sets. Since the observed values are the same across the m data sets, and the missing values are filled in with different imputations, this method reflects uncertainty levels. The variation across the cells could be small when the model predicts well or it may be larger, or asymmetric to reflect whatever knowledge and level of certainty is available about the missing information (King et al., 2001). Next, the m datasets are analyzed, and quantities of interest are estimated, we can apply the statistical method that we would have used if there were no missing values, in each of the m data sets (King et al., 2001). Lastly, the m separate point estimates are combined into one according to the "Rubin combination rules" (Rubin, 1987).

Rubin's combination rules:

First, compute a Quantity of Interest (denoted as Q), which could represent a univariate mean, regression coefficient, predicted probability, or first difference, for each dataset (indexed as $j = 1, 2, \dots, m$). The overall point estimate of \bar{q} is then obtained by averaging the m individual estimates, q_j :

$$\bar{q} = \frac{1}{m} \sum_{j=1}^m q_j$$

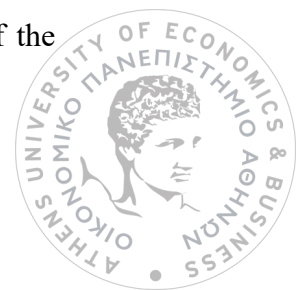


Let $SE(q_j)$ denote the estimated standard error of q_j from the data set j , and $S_q^2 = \sum_{j=1}^m \frac{(q_j - \bar{q})^2}{(m-1)}$ the sample variance across the m point estimates. The variance of the point estimate obtained through multiple imputation is computed as the average of the estimated variances calculated within each completed dataset, augmented by the sample variance in the point estimates across all datasets. Additionally, this value is adjusted by a correction factor to account for bias arising from the finite number of datasets ($m < \infty$).

$$SE(q)^2 = \frac{1}{m} \sum_{j=1}^m SE(q_j)^2 + S_q^2 \left(1 + \frac{1}{m}\right)$$

As discussed by Lall (2016) the recommendation, and the default setting in [AMELIA II](#) (King et al., 2011) is to conduct no more than five imputations. This recommendation, as Lall (2016) explains, is based on Rubin's (1987) formula for the relative efficiency of a parameter estimate derived from m imputations compared to one from infinitely many imputations is given by $(1 + \gamma/m)^{-1}$, where γ represents the "fraction of missing information." This quantity, although intricate, generally reflects the extent of information loss concerning the parameter due to missing data (Lall, 2016). This formula suggests that the efficiency of an estimate when $m = 5$ consistently approaches that of one when $m = \infty$ (Lall, 2016). It's important to mention that recent research discussed and analyzed by Lall (2016) indicate that m should be approximately equal to the percentage of incomplete observations in the data, in order to avoid unwanted levels of statistical power and precision (Bodner ; Graham, Olchowski & Gilreath; White, Royston & Wood as cited in Lall 2016). Lall (2016) argues that this rule has the risk that as the number of variables in the imputation model increases, the percentage of incomplete observations rapidly falls to zero. This is why Lall (2016) adopts a different version of the rule where m is equal to the average missing data rate of all variables in the imputation model.

In the multiple imputation process, three fundamental stages are involved. Initially, m values are assigned to each missing cell, introducing variability in these values to account for uncertainties in the imputation model. These imputed values are drawn independently from a posterior distribution of the missing data, given the observed data. Typically, this distribution is derived from a parametric model if the



complete data adhere to a joint probability distribution with unknown parameters. Commonly, this model is based on a multivariate normal distribution, despite the acknowledgment that real-world data may not strictly adhere to multivariate normality (Lall, 2016). Nevertheless, this modeling approach has demonstrated effective performance even in the presence of deviations from normality (Rubin & Schenker, Schafer as cited in King et al., 2001). The fit of the model can be improved by transformation of the variables and other techniques (King et al., 2001). The authors explain that if the data are MAR, then the multiple imputations from the normal model will nearly always surpass current practice.

The likelihood function for complete data is expressed by King et al. (2001) as follows:

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^n N(D_i | \mu, \Sigma).$$

- D is the data matrix which includes the dependent variable Y and the independent variables X such as $D = \{Y, X\}$.
- D_i , $i=1, \dots, n$: the vector of values of the p variables (dependent Y_i and explanatory X_i) which if all of them were observed they would follow the Normal distribution, with parameters: mean vector μ and variance matrix Σ . The off – diagonal elements of Σ allow variables within D to depend on one another.

They continue by forming the observed data likelihood, after assuming that data are MAR. The observed data likelihood given that the marginal densities are normal is the following (King et al., 2001):

$$L(\mu, \Sigma | D_{obs}) = \prod_{i=1}^n N(D_{i,obs} | \mu_{i,obs}, \Sigma_{i,obs}).$$

- D_{obs} is the observed portion of D and D_{mis} is the missing portion of D : $D = \{D_{obs}, D_{mis}\}$.
- $D_{i,obs}$: the observed elements of row i of D
- $\mu_{i,obs}$: the corresponding subvector of μ .
- $\Sigma_{i,obs}$: the corresponding submatrix of Σ .

The multivariate normal specification assumes that missing values are imputed linearly, similar to predicting values in a regression model. For instance, if

- \tilde{D}_{ij} : the imputed value for observation i and variable j .



- $D_{i,-j}$: the vector of values of all observed variables in row i , except for variable j .

To create a simulated value we use a coefficient β derived from a regression of D_j on the other variables in the dataset. This coefficient β can be directly calculated from the mean vector (μ) and the variance matrix (Σ). We then use this coefficient to predict the missing value according to the equation:

$$\tilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i$$

Here, $\tilde{\beta}$ represent a random draw from the posterior distribution of β and $\tilde{\varepsilon}_i$ is a random error term. The symbol " \sim " indicates that these values are randomly sampled from the appropriate posterior.

In essence, that random draws of \tilde{D}_{ij} are linear functions of the other observed variables ($D_{i,-j}$), the uncertainty in the estimated coefficient ($\tilde{\beta}$), and the fundamental uncertainty ($\tilde{\varepsilon}_i$). Even with an infinite sample, there would still be some uncertainty due to the variability in the real world, represented by ε_i . The main computational challenge is taking random draws from the posterior of μ and Σ (King et al., 2001).

5.2.1 Handling Categorical variables

The equation mentioned above can be applied to generate imputations for categorical variables by rounding to the nearest valid integer (Schafer as cited in King et al., 2001). A slightly more effective method involves drawing from a multinomial or other suitable discrete distribution with the mean equivalent to the normal imputation. For instance, to impute a 0/1 variable, we take a Bernoulli draw with the mean equal to the imputation (truncated to [0,1] if necessary). In other words, we impute a 1 with a probability equal to the continuous imputation, otherwise 0 (King et al., 2001). Additionally, we can deal with categorical variables by drawing imputations from conditional distributions or from observed values of similar units. Research has shown that these techniques give similar results as [AMELIA II](#) (Lall, 2016). King et al. (2001), underline that many variables that come from political and social science surveys are ordinal variables with more than four different values. They explain that these types of variables are reasonably well approximated by the normal model during the imputation process.



An important issue is how to handle missing values in ordinal and nominal data. The [AMELIA](#) package provides tools for imputing missing values in such data types. In many statistical studies, researchers often treat independent ordinal variables, including dichotomous ones, as if they were continuous. If the chosen analytical model follows this approach, there is no additional requirement for the imputation model (King et al., 2011). Users are encouraged to permit [AMELIA](#) to impute non-integer values for any missing data and incorporate these non-integer values into their analysis. While this approach may be logical in certain cases, at times it may seem counterintuitive. In the case of nominal variables, they must be treated differently, any multinomial variables must be specified to [AMELIA](#) (King et al., 2011).

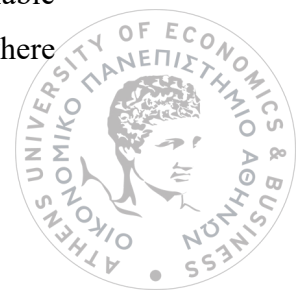
5.2.2. Computational algorithms (King et al., 2011)

To implement multiple imputation, we utilized the [AMELIA](#) package in R (AMELIA I: King et al., 2001, AMELIA II: King et al., 2011). AMELIA assumes that our data are jointly distributed as multivariate normal. AMELIA II, the latest version, employs an Expectation Maximization algorithm with bootstrapping (EMB), while AMELIA I used an Expectation Maximization (EM) algorithm (Dempster, Laird & Rubin 1977).

We will mention three computational algorithms which are proposed by King et al. (2001) in *Analyzing Incomplete Political Science Data: An alternative algorithm for Multiple Imputation*. These four computational algorithms, as the authors underline, solve the problem of computing the observed data likelihood and taking randoms draws from it, which is computationally impossible with classical methods.

The first one is Imputation-Posterior (IP) allowing us to generate random simulations from the multivariate normal posterior distribution of observed data posterior $P(D_{mis}|D_{obs})$, this process involves two iterative steps (King et al., 2001). It can be adapted to numerous specialized models, but it is slow and hard to use.

On the contrary, the Expectation-Maximization (EM) algorithm is characterized by its speed, deterministic convergence, and the consistent increase of the objective function with each iteration. The EM algorithm, introduced by Rubin et al. (1977), is a powerful statistical technique designed for handling incomplete or missing data in the context of maximum likelihood estimation. The EM algorithm consists of two main steps: the Expectation step, where missing data is imputed based on available information and current parameter estimates, and the Maximization step, where



parameters of the statistical model are updated by maximizing the expected log-likelihood, considering both observed and imputed data. This iterative process continues until convergence, providing robust estimates of parameters even when dealing with incomplete datasets. The EM has also disadvantages. For example, it might get stuck in local maxima of the likelihood function instead of finding the global maximum. In addition, the EM provides only the maximum likelihood estimates for parameters and imputations and not the full distribution. This can be challenging because the estimation uncertainty is ignored. When we use the EM for multiple imputation, it assumes that parameter estimates are known with certainty, and it ignores the uncertainty in the estimates.

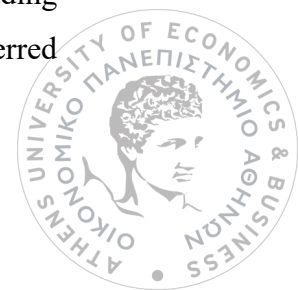
As denoted by King et al. (2001) they begin with EM and then they add back in estimation uncertainty, so they get draws from the correct posterior distribution of D_{mis} . They call this algorithm EMs (EM with sampling), they explain that it is very fast, it produces independent imputations, it converges nonstochastically and it works well in large samples but it performs poorly with small samples or many parameters.

This is why they propose the EMis algorithm (Expectation-Maximization with importance resampling), which draws samples from the identical posterior distribution as IP but with significantly enhanced speed. For this particular model, there seem to be no challenges related to convergence or independence (King et al., 2001). The authors note that, EMis follows the same steps as EMs except draws of θ from its asymptotic distribution are treated only as first approximations to the true (finite sample) posterior.

The Expectation Maximization algorithm with bootstrapping (EMB algorithm) as outlined by King et al. (2011) in their paper *Amelia II: A Program for Missing Data* combines the classic EM algorithm with a bootstrap strategy to draw samples from the posterior distribution. For each iteration, we utilize bootstrapping on the dataset to simulate estimation uncertainty. Then, employing the EM algorithm, we determine the mode of the posterior distribution for the bootstrapped data, thus incorporating fundamental uncertainty, as described by the authors.

5.2.3. Why choose multiple imputation?

Multiple imputation surpasses listwise deletion in efficiency for two primary reasons. Firstly, it employs data from incomplete observations rather than discarding them. Secondly, it enables analysts to enhance the imputation model by including additional information through variables not directly involved in the analysis, referred



to as "auxiliary variables." This utilization of auxiliary variables contributes to the overall effectiveness of multiple imputation compared to the less efficient strategy of listwise deletion (Lall, 2016).

In essence, the objective of multiple imputation is to retain essential characteristics of the available data, such as means, variances, and covariances. Simultaneously, it aims to account for the uncertainty associated with predicting missing data, ensuring a comprehensive representation of the underlying data structure (Lall, 2016). Despite the potential for bias when there are weak associations between variables and the occurrence of missing data, multiple imputation remains the preferred strategy over listwise deletion.

Concluding the above unit, we explained the multiple imputation and in particular its implementation by the [AMELIA](#) package in R. King et al. (2001;2011) first utilized the EM and the EMis as computational algorithms and then in the last version they introduce the EMB. Moreover, we explain how categorical variables should be treated when performing multiple imputation with [AMELIA](#). Lastly, we introduce some advantages of multiple imputation compared to listwise deletion.

5.3. Logistic regression and Variable selection

Our goal in this part is to explore if the data are MAR with respect to the missing values in the voting question and identify which variables are related with the nonresponse in the vote choice question. Additionally, we aim to construct a model capable to predict vote nonresponse using the rest covariates. After imputing for the missing values in the independent variables, we employed logistic regression analysis to examine the determinants of nonresponse to the question regarding voting behavior in the last national election ($Y = 0$: response, $Y = 1$: nonresponse). The logistic regression model is expressed as:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Where:

- $P(Y = 1)$ represents the probability of nonresponse.
- $\text{logit}(\cdot)$ is the log-odds function.
- β_0 is the intercept term.



- $\beta_1, \beta_2, \dots, \beta_p$ are the coefficients corresponding to predictor variables X_1, X_2, \dots, X_p respectively.

The predictor variables included in the model were selected based on variable selection techniques. These variables encompassed socio-demographic characteristics, attitudes, and other relevant factors that could potentially influence nonresponse behavior.

5.3.1. LASSO and GROUP LASSO

Since we deal with a dataset characterized by large number of potential predictors, we need to reduce the number of predictors and find the variables that influence nonresponse. We decide to perform variable selection with LASSO (Least Absolute Shrinkage and Selection Operation) (Tibshirani, 1996). This technique extends the classical regression by introducing a penalty term that encourages the model to favor sparse solutions, effectively promoting variable selection. Moreover, due to the dimension of the data, we may face challenges of multicollinearity and overfitting. In such scenarios, regularization techniques come to forefront. For LASSO we need to minimize the following function:

$$\min_{\beta_0, \beta} \{ \sum_{i=1}^n (y_i - \beta_0 - x_i^T \beta)^2 \} \text{ subject to } \sum_{j=1}^p |\beta_j| \leq \lambda$$

Where:

- y_i is the outcome.
- $x_i := (x_1, \dots, x_p)_i$ is the vector of covariates for the i^{th} case.
- $\beta := (\beta_1, \dots, \beta_p)$ is the coefficient vector and β_0 the constant coefficient.
- λ is the parameter that determines the regularization degree.

In essence, LASSO introduces sparsity to the model, driving some of the regression coefficients to exact zero. This unique property renders LASSO a powerful tool for automatic variable selection, particularly valuable in situations where the dataset contains a multitude of potential predictors. The selection of the regularization parameter λ is a critical aspect of LASSO, and its optimal value is often determined through techniques like cross-validation.



However, in the case of categorical covariates, LASSO may fail, as it only selects individual dummy variables instead of factors. We decided to explore the group-lasso (Yuan & Lin, 2005), which is a generalization of LASSO for groupwise variable selection. Group – Lasso can be applied to select pre-defined groups of predictors. In our case, d predictors are divided into 124 groups (the number of independent variables). The group-lasso estimators are obtained by minimizing:

$$\sum_{i=1}^{2327} \left(y_i - \sum_{j=1}^{124} X_j \beta_j \right)^2 + \lambda \sum_{j=1}^{124} \|\beta_j\| \kappa_j$$

where λ is the regularization parameter. The entire vector β_{κ} will either be entirely zero, or all its elements will be nonzero. Hence, the penalty is not applied to each variable separately but to batches; either all variables shrink towards zero, or none. We implement group-lasso using a fast unified algorithm for solving group-lasso penalized learning problems (Yang & Zou, 2015). Nevertheless, the group-lasso is computationally more challenging than the LASSO. The least angle regression (LARS) algorithm (Efron et al., 2004), is not proper for group-lasso penalized least squares since its solution paths are not piecewise linear.

We will implement group-lasso in R using the [gglasso](#) package (Yang & Zou, 2015) which contains the function for fitting the group-lasso logistic regression. Yang and Zou (2015) suggest a simple unified algorithm, *groupwise – majorization – descent (GMD)*, for solving the general group-lasso learning problems. The above can be done under the condition that the loss function satisfies a Quadratic Majorization (QM) condition.

The QM condition:

We present the definition given by Yang and Zou (2015) in their work *A fast unified algorithm for solving group-lasso penalize learning problems*.

Definition

The loss function Φ is said to satisfy the quadratic majorization (QM) condition, if and only if the following two assumptions hold:

- i. $L(\beta | D)$ is differentiable as a function of β i.e., $\nabla L(\beta | D)$ exists everywhere.
- ii. There exists a $p \times p$ matrix H , which may only depend on the data D , such that for all β, β^* ,



$$L(\beta|D) \leq L(\beta^*|D) + (\beta - \beta^*)^T \nabla L(\beta^*|D) + \frac{1}{2} (\beta - \beta^*)^T H(\beta - \beta^*).$$

As Yang and Zou (2015) explain the QM condition is verified for the logistic regression that we care about.

GMD algorithm

Algorithm 1 The GMD algorithm for general group-lasso learning

1. For $k=1, \dots, K$, compute γ_k , the largest eigenvalue of $\mathbf{H}^{(k)}$.
2. Initialize β .
3. Repeat the following cyclic groupwise updates until convergence:

For $k=1, \dots, K$ do step (3.1)-(3.3)

3.1 Compute $U(\tilde{\beta}) = -\nabla L(\tilde{\beta}|D)$.

3.2 Compute

$$\tilde{\beta}_{(new)}^{(k)} = \frac{1}{n} (U^{(k)} + \gamma_k \tilde{\beta}^{(k)}) \left(1 - \frac{\lambda \omega_k}{\|U^{(k)} + \gamma_k \tilde{\beta}^{(k)}\|_2} \right)$$

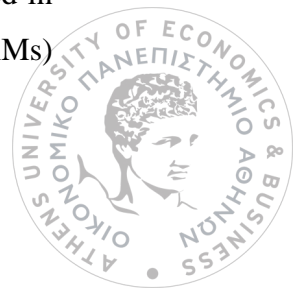
3.3 Set $\tilde{\beta}^{(k)} = \tilde{\beta}_{(new)}^{(k)}$.

The above method is useful for implementing logistic regression with Group Lasso. Firstly, we standardize the data in advance, ensuring each numeric variable has zero mean and unit sample variance. We fit a sparse additive logistic regression model, utilizing the group lasso penalty to select important additive components. As mentioned earlier, our data contain both categorical and numeric variables. For any categorical variable with M levels of measurement, we recode it into $M - 1$ dummy variables and treat these dummy variables as a group. For each continuous variable, we use five B-spline basis functions to represent its effect in the additive model, treating those five basis functions as a group (Yang & Zou, 2015). We include the intercept in the model, and the intercept forms a group.

We discussed the LASSO regularization as a variable selection technique. We explained that LASSO may perform poorly when having categorical predictors. We employed the group-lasso method. Lastly, we presented the QM condition and the GMD algorithm for implementing group-lasso.

Generalized Additive Models (GAMs)

We provide a brief presentation of the additive models since they are used in order to implement the group-lasso shrinkage. Generalized Additive Models (GAMs)



provide a framework for modeling non-linear data while preserving interpretability. They relax the constraint of assuming a simple weighted sum relationship, allowing the outcome to be expressed as a sum of arbitrary functions for each feature. Instead of using beta coefficients as in Linear Regression, GAMs employ flexible functions known as splines to capture non-linear relationships. Splines are intricate mathematical functions that enable the modeling of complex, non-linear patterns for each feature. By combining the contributions of multiple splines, GAMs create a highly flexible model that maintains some of the interpretability characteristic of linear regression.

5.3.2. Stepwise variable selection via AIC

After applying Group Lasso, the model might still include some redundant or non-significant variables. This is why we will perform stepwise variable selection using AIC that can lead to a more parsimonious model by removing variables that do not significantly improve model fit, beyond the shrinkage imposed by Group Lasso.

We will use the Akaike Information Criterion proposed by Akaike (1971;1974):

$$AIC = -2 \log L(\hat{b}) + 2k$$

- k is the number of parameters in the model
- $\log L(\hat{b})$ is the natural algorithm of the likelihood function of the model.

We start from a given model and in every step, we evaluate which variable to include in the model. We always select the one with the minimum AIC value. After adding the optimal variable, we evaluate if any of the included should be removed. In each step we decide according to the maximum or the minimum criterion value. We stop when no other improvement can be achieved in our model. We prefer the stepwise variable selection compared to backward and forward because of the double evaluation.

5.4. Logistic Regression Diagnostics

We will compare the models and evaluate their performance to identify the optimal one. To assess model performance, we conduct 5-fold cross-validation predictions and generate ROC curves for each fold.



5.4.1. Receiver Operation Characteristic (ROC) Curve

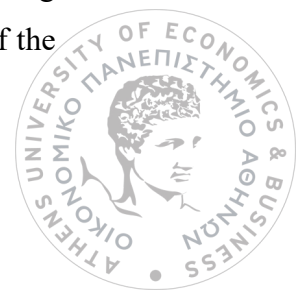
The ROC curve, is an analytical method, represented as a graph, that is used to evaluate the performance of a binary diagnostic classification method (Nahm, 2022). ROC graphs are two-dimensional graphs in which True Positive rate is plotted on the Y-axis and False Positive rate is plotted on the X-axis (Fawcett, 2006). The diagonal line $y = x$ represents the random guess of a class. The Area Under the Curve (AUC), is a useful measure in order to compare classifiers. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0 (Fawcett, 2006). However, because the diagonal line between (0, 0) and (1, 1) represent the random guess, which has an area of 0.5, no realistic classifier should have an AUC less than 0.5 (Fawcett, 2006).

5.4.2. K- fold cross validation

Cross validation is a method we use to estimate the predictive performance of one or more candidate models. This method consists of data splitting, then fitting the model to a subset of the data (the training set) and then compare the model's performance on the remaining data (the test set) (Yates et al., 2023). The method can be described as training and testing the model. As Yates et al. (2023) point out there is a variety of cross validation methods. A method differs from the others for its data-splitting scheme. We will use a famous approach, the k-fold cross validation. The splitting of the data in this method is done by dividing them in k equal-sized subsets, the 'folds', which lead to k distinct pairs of training and testing set. We train the model using the $k - 1$ subsets, which represent the training set. Next, we test the model using the remaining subset, which is the validation/testing set. The performance is finally measured, and we repeat the procedure until each of the k subsets is used as testing set. This method can be applied once, as explained above, for a single split (ordinary k-fold) or we can repeat it for different splits (repeated k-fold). We choose the first approach (Yates et al., 2023).

5.5. Clustering Techniques with Missing Data

We aim to cluster the nonresponders to identify patterns which can lead us to understand better the profile of those who retain their vote choice. The main problem we faced was the missing data in all the independent variables, we needed a clustering technique which could be applied to a data set with missing values. This is one of the



reasons we chose to use the [BayesBinMix](#) package (Papastamoulis & Rattray, 2017) for model-based clustering of multivariate binary data. The authors explain why they believe that the Bayesian approach is more suitable for clustering binary data:

1. **Suitable for handling Missing Data:** The package can process datasets with missing values without requiring imputation or listwise deletion which can lead to loss of valuable information.
2. **Bayesian Approach:** Unlike the Expectation-Maximization algorithm, which may lead to local maxima and convergence issues, under the Bayesian approach we can put a prior distribution on the number of clusters, the model parameters and then (approximately) sample from the joint posterior distribution using Markov Chain Monte Carlo (MCMC) algorithms.
3. **Number of Clusters:** The Bayesian framework allows for the estimation of the number of clusters directly from the data, using prior distributions and posterior sampling. It uses the MCMC in order to estimate the posterior distribution of the number of clusters and the model parameters. In addition, it can produce a rapidly mixing MCMC sample by running parallel heated chains which can switch states. Finally, the package analyses the sample generated by the MCMC and gives as result meaningful posterior mean estimates after dealing with label switching (Papastamoulis & Rattray, 2017).

The authors assume that the observed data $x = (x_1, \dots, x_n)$ consists of binary vectors, where $x_i = (x_{i1}, \dots, x_{id})$, $d > 1$, for $i = 1, \dots, n$. Each observation x_i has been generated from one of K clusters, each represented by a multivariate Bernoulli distribution. a mixture of independent Bernoulli distributions. The mixture model is the following:

$$x_i \sim \sum_{k=1}^K p_k \prod_{j=1}^d f(x_{ij}, \theta_{kj}) = \sum_{k=1}^K p_k \prod_{j=1}^d \theta_{kj}^{x_{ij}} (1 - \theta_{kj})^{1-x_{ij}} I_{\{0,1\}}(x_{ij})$$

Independently for $i = 1, \dots, n$

- $\theta_{kj} \in \Theta = (0,1)$: the probability of success for the k – th cluster and the j – th response for $k = 1, \dots, K$; $j = 1, \dots, d$.
- $p = (p_1, \dots, p_k) \in P_{K-1} = \{p_k; k = 1, \dots, K - 1 : 0 \leq p_k \leq 1; 0 \leq p_k = 1 - \sum_{k=1}^{K-1} p_k\}$: the vector of mixture weights



- $I_A(\cdot)$: the indicator function of a measurable subset A

Next Papastamoulis and Rattray (2017), explain the prior assumptions of the model.

The prior assumptions that are imposed are the following:

- Number of clusters (K): $K \sim Discrete \{1, \dots, K_{max}\}$, the discrete distribution could be either a Uniform or a Poisson with mean $\lambda=1$ truncated on the set $\{1, \dots, K_{max}\}^2$.
- Mixture Weight (p): $p|K \sim Dirichlet(\gamma_1, \dots, \gamma_k)$, we set $\gamma_1 = \dots = \gamma_k = \gamma > 0$.
- Bernoulli Parameters (θ): $\theta_{jk}|K \sim Beta(\alpha, \beta)$

they are independent for $k = 1, \dots, K$; $j = 1, \dots, d$.

The joint probability density function of the model is:

$$f(x, K, z, p, \theta) = f(x|K, z, \theta)f(z|K, p)f(p|K)f(\theta|K)f(K).$$

The *BayesBinMix* package allows to jointly estimate the number of cluster and the model parameters using MCMC sampling. To be more specific, they use the Metropolis-coupled MCMC (MC³) sampler. MC³ which also utilizes parallel heated chains. This sampler can make convergence to the target posterior distribution faster.

Finally, *BayesBinMix* package also handles label switching. Label Switching is a common problem which occurs when the labels of the clusters get confused during the sampling process. The authors solve this problem by using different algorithms.

5.6. Conclusion

In this chapter, we explained thoroughly the methods we used to assess the characteristics of nonresponders and validate the assumption that our data are MAR. We started with the imputation of missing data, then we explained the screening techniques we will use and the variable selection to fit the optimal model. Finally, we presented the *BayesBinMix* package that we will use to cluster the nonresponders.





CHAPTER 6

Implementation – Results

6.1. Introduction

In the upcoming chapter, we will outline the outcomes derived from our analysis. Firstly, we will detail the findings of our imputation process and diagnostic assessments. Subsequently, we will delve into the execution of LASSO and Group-Lasso regularization techniques. Following this, we'll demonstrate the outcomes of stepwise selection and determine the most suitable model using ROC curves. ROC curves will be used to evaluate the final model performance. Lastly, we will interpret the coefficients of the optimal model and implement clustering to create subgroups within the nonresponders.

6.2. Dealing with missing data

As we already mentioned all the independent variables contain missing values. We decided to implement multiple imputation to obtain a new data set without missing values. We will proceed with our analysis using the new implemented data set. In many statistical studies, researchers often handle independent ordinal variables as if they were continuous. As King et al. (2011) discuss, if the analysis model falls into this category, then no additional steps are necessary for the imputation model. They recommend users to permit AMELIA to impute non-integer values for any missing data and utilize these non-integer values in their analysis.

We decided to conduct two imputation processes. One involved treating the answers to the questions as distinct categories, acknowledging that ordinal variables cannot be imputed as continuous. The second imputation assumed that ordinal variables would be treated as continuous covariates in our model.

Multiple imputation was performed using AMELIA, where we specified the categorical variables and in the first approach, we also specified the ordinal variables. The variable *nwspol*, which exhibited skewness, was log-transformed. This process resulted in 5 imputed datasets with no missing values. We used seed '123' in both imputation processes for reproducibility reasons. Next, we tested the imputation performance by implementing several diagnostics. According to the multiple imputation diagnostics, there are no foolproof tests of the assumptions of the imputation procedures, we cannot test unobserved values for agreement with an unknown true



distribution (Abayomi et al., 2008). Moreover, we created the densities of the numeric observed and imputed data, nevertheless differences in distribution between the imputed and the observed data do not necessarily indicate violations of the missingness assumptions or problems with the imputation model. Some deviations between observed and missing values can be expected under MAR (Abayomi et al., 2008). Lastly, to combine these datasets, we calculated the mean value for each cell across the 5 imputed sets for numeric variables and the mode for each cell across the 5 imputed sets for categorical variables. We decided to create graphs for the distribution of the data before and after the multiple imputation.

6.3. Variable selection using LASSO and Group – Lasso.

After completing the imputation process for missing values, we now possess two datasets: one where ordinal covariates are considered continuous, and another where ordinal covariates are treated as factors. Our objective is to determine the variables that influence the occurrence of nonresponse in the question concerning the vote choice. To accomplish this, we will employ a logistic regression model with $Y=0$ denoting response and $Y=1$ indicating nonresponse in this question. Initially, we will utilize LASSO and group-lasso methods to identify significant covariates and create a model able to predict nonresponse.

Firstly, we consider ordinal covariates as continuous, and we perform LASSO regularization. The optimal lambda value is determined through 10-fold cross-validation, and we set seed ‘123’ to be able to recreate the results. The optimal lambda value is λ_{min} which is the value of λ that gives minimum mean cross-validated error. On the other hand, λ_{1se} is the value of λ that gives the most regularized model such that the cross-validated error is within one standard error of the minimum (see Table 6.1). We created Figure 6.1 to evaluate and represent the cross-validation curve (depicted as a red dotted line) accompanied by upper and lower standard deviation curves across the λ sequence, represented as error bars. Additionally, two specific values along the λ sequence are highlighted by vertical dotted lines (Hastie et al., 2023). Subsequently, a LASSO model is fitted with $\lambda = 0.0053$, retaining **77 non-zero covariates**.



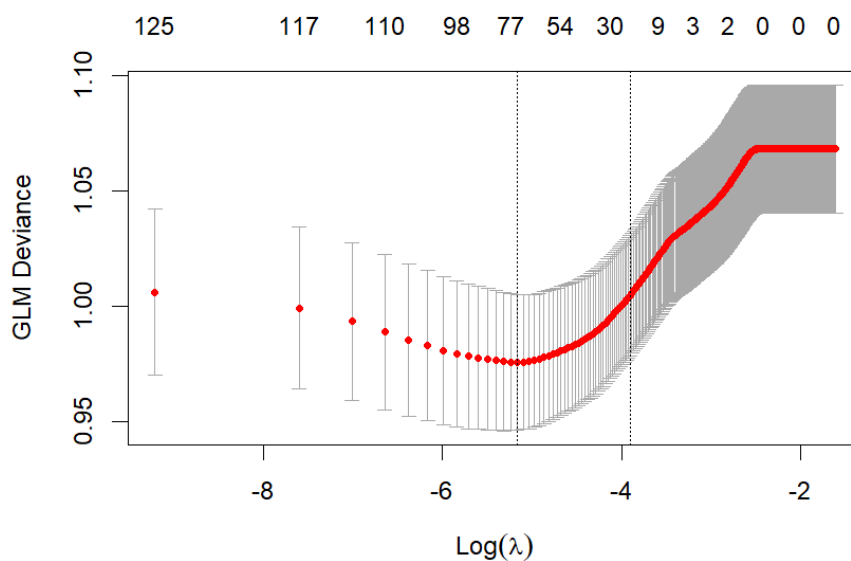


Figure 6.1: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for LASSO (assuming ordinal covariates as continuous)

λ_{min}	0.0053
λ_{1se}	0.019

Table 6.1: The optimal lambda value λ_{min} and the λ_{1se} for LASSO.

Transitioning to group-lasso regularization, the optimal λ value is selected again by performing a 10-fold cross-validation. We set as seed '123' for reproducibility reasons. The cross-validation curve is illustrated in Figure 6.2 where the vertical dotted lines show the values of λ_{min} and λ_{1se} . After fitting the group-lasso model with $\lambda_{min} = 0.0007$ (see Table 6.2), we keep **79 non- zero covariates**.

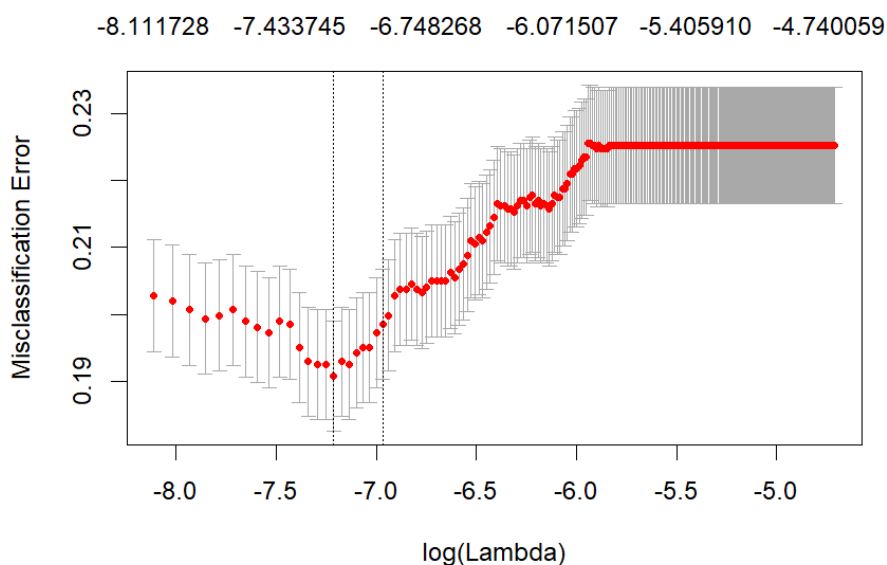


Figure 6.2: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for Group-Lasso (assuming ordinal covariates as continuous)

λ_{min}	0.0007
λ_{1se}	0.0009

Table 6.2: The optimal lambda value λ_{min} and the λ_{1se} for Group-Lasso.

Extending our analysis to treat ordinal covariates as factors in the model, first LASSO is applied, the optimal λ value is selected again by performing a 10-fold cross-validation, we set as seed '123' for reproducibility reasons. We can observe the graphical representation of the cross-validation curve, in Figure 6.3 below. The figure reveals that $\lambda_{min} = 0.0053$. We fit the LASSO model with an optimal $\lambda_{min} = 0.0053$ and **73 covariates are retained**.

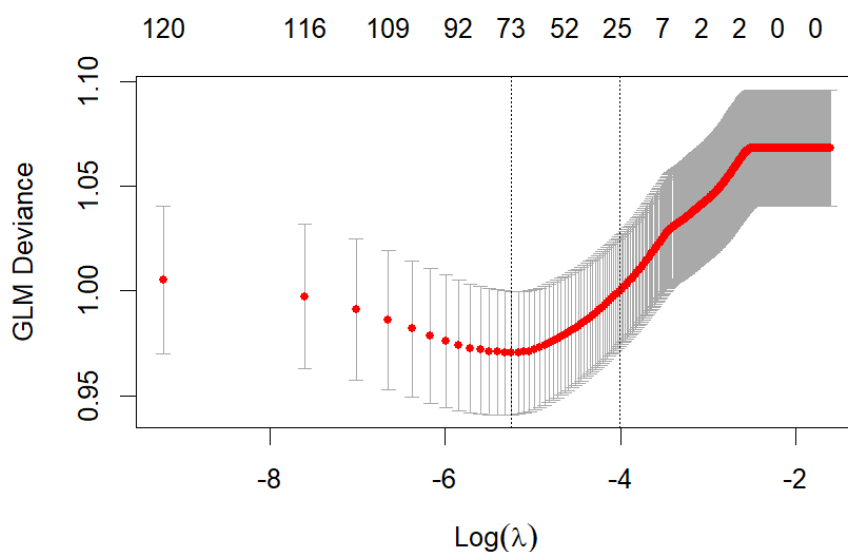


Figure 6.3: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for LASSO (assuming ordinal covariates as factors)

λ_{min}	0.0053
λ_{1se}	0.02

Table 6.3: The optimal lambda value λ_{min} and the λ_{1se} for LASSO.

Group-lasso regularization is now applied in the second dataset, the optimal λ value is selected again by performing a 10-fold cross-validation, we set as seed '123' for reproducibility reasons. We can observe the graphical representation of the cross-validation curve, in Figure 6.4 below. The figure reveals that $\lambda_{min} = 0.0009$. We fit the group-lasso model with an optimal $\lambda_{min} = 0.0009$ and **89 covariates are retained**.

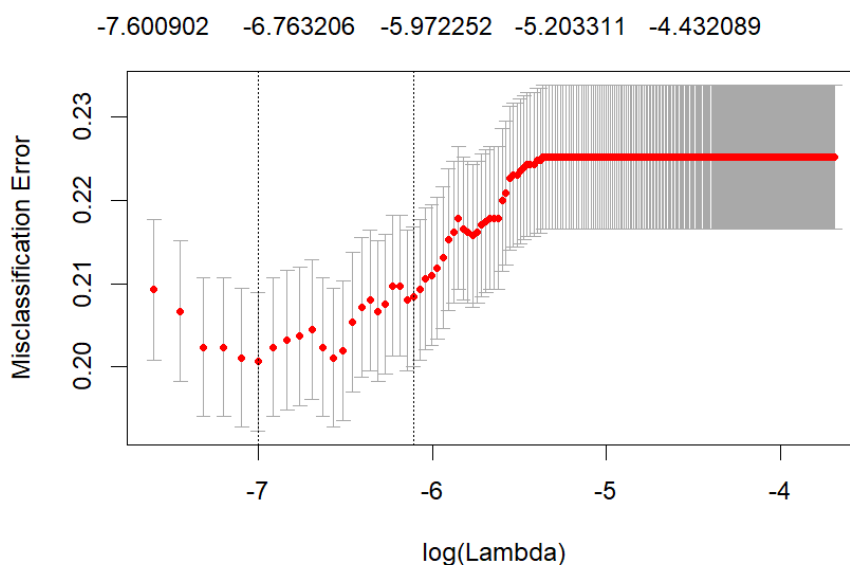


Figure 6.4: 10- fold cross validation plot $\text{Log}(\lambda)$ vs Deviance for Group-Lasso (assuming ordinal covariates as factors)

λ_{min}	0.0009
λ_{1se}	0.002

Table 6.4: The optimal lambda value λ_{min} and the λ_{1se} for Group-Lasso.

6.4. Stepwise variable selection using AIC.

Following this we performed stepwise variable selection using AIC criterium to further reduce model's dimension. We performed variable selection using the common stepwise in both directions and the AIC as criterion. The best model is chosen by seeing which has the lowest AIC value. We start with the full model (the one with all the covariates we kept after lasso). The goal is to obtain a more parsimonious model.

For the first model, where ordinal covariates were treated as continuous and LASSO was used for regularization, after stepwise variable selection we end up with a model with **43 covariates** and an AIC=2147.6 (Table 6.5, Appendix).

For the second model, where once more ordinal covariates were treated as continuous, we implemented group-lasso, after stepwise variable selection we have a model with **41 independent variables** and an AIC=2157.1 (Table 6.6, Appendix).

For the third model, where ordinal covariates were treated as factors and we implement group-lasso, after stepwise variable selection we have a model with 37 covariates, we discard one coefficient because of the large standard error. A large standard error indicates that this covariate doesn't have values in all its categories. The final model has **36 covariates** and an AIC=2086.3 (Table 6.7, Appendix).

For the fourth model, where ordinal covariates were treated as factors and we implement group-lasso, after stepwise variable selection we have a model with 39 covariates, we discard one coefficient because of the large standard error. A large standard error indicates that this covariate doesn't have values in all its categories. The final model has **38 covariates** and an AIC=2085.6 (Table 6.8, Appendix).

6.5. Evaluating Model Performance with ROC Curves

To assess the performance of our logistic regression models, we employ ROC (Receiver Operating Characteristic) curves and the AUC (Area Under the Curve) as key metrics. We implement 5-fold cross-validation, and we compute the ROC curve for each fold. For all models the AUC values are below 0.8. According to Nahm (2021), for any diagnostic technique to be meaningful, the AUC must be greater than 0.5, and in general, it must be greater than 0.8 to be considered acceptable. The fact that our models' AUC values are over 0.5 and below 0.8 indicates fair diagnostic accuracy but suggests room for improvement. Below we present the ROC curves for each model after 5-fold cross-validation (Figures 6.5-6.8). For each of the 5 folds we compute and represent the ROC curve. All models perform better than a random prediction but, the model with the best performance seems to be *Model 4* which assumes ordinal covariates as factors.



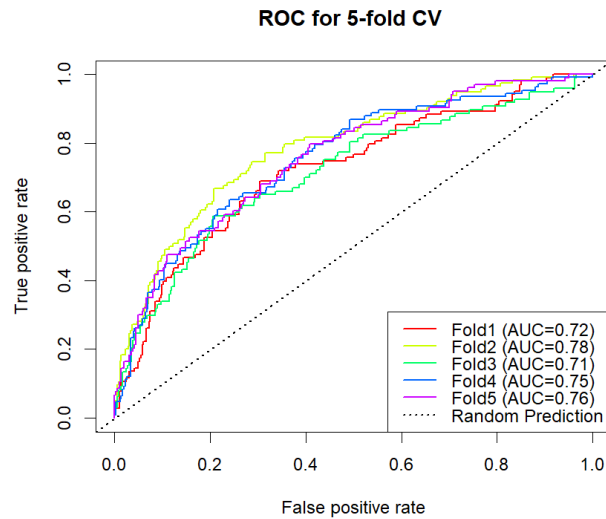


Figure 6.5: ROC curves for 5-fold cross validation (assuming ordinal covariates as continuous). Model 1 after lasso and stepwise variable selection.

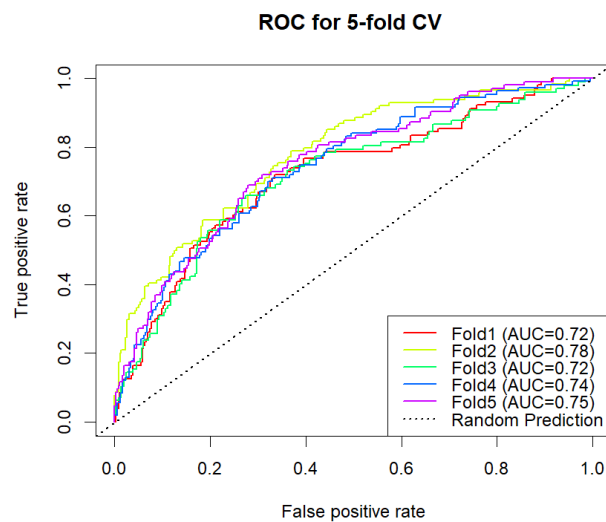


Figure 6.6: ROC curves for 5-fold cross validation (assuming ordinal covariates as continuous). Model 2 after group – lasso and stepwise variable selection.

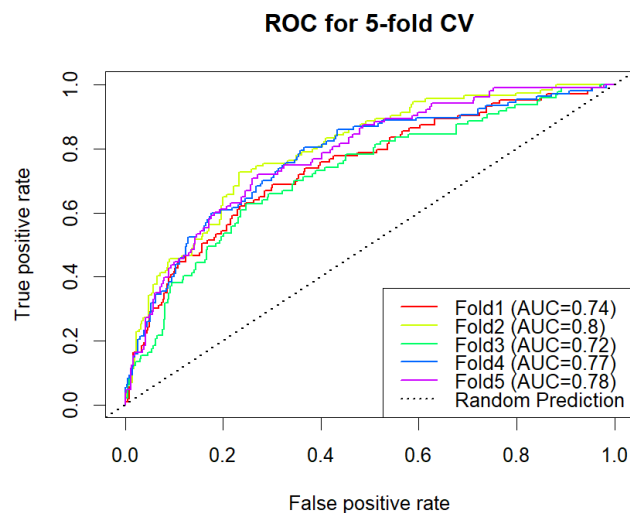


Figure 6.7: ROC curves for 5-fold cross validation (assuming ordinal covariates as factors). Model 3 after lasso and stepwise variable selection.

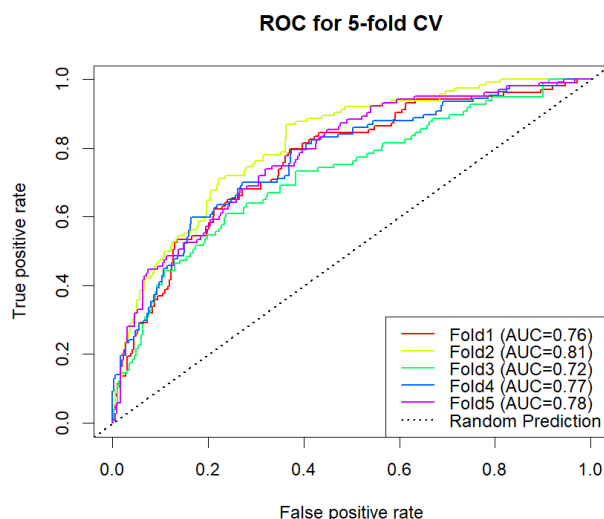


Figure 6.8: ROC curves for 5-fold cross validation (assuming ordinal covariates as factors). Model 4 after group – lasso and stepwise variable selection.

Notably, the model employing ordinal variables as factors exhibits the best AIC performance among the three. The comparison of AIC values supports the selection of the model treating ordinal variables as factors as the optimal choice for estimating nonresponse.

6.6. Profiling Non – Responders: Insights from the Optimal Model

The logistic model provides insights into the characteristics of nonresponders to the voting choice question in the last national elections based on the exponentiated

coefficients. Here are the key insights based on the exponentiated coefficients from *Model 4*: The intercept, with an exponentiated coefficient of 0.37, suggests the baseline probability of nonresponse.

Characteristics that reduce nonresponse probability (odd ratio < 1)
Higher trust in the legal system, parliament, and politicians.
Belief in the importance of punishing governing parties for poor performance
Belief that national elections are free and fair, that governing parties are punished when they have done a bad job, and that parties offer clear alternatives
Not identifying with far right or center ideologies.
Valuing understanding different perspectives.

Table 6.9: Characteristics of participants that reduce nonresponse probability.

Characteristics that increase nonresponse probability (odd ratio > 1)
Satisfaction with the government and health services.
Belief that minority rights are protected, that the government protects citizens from poverty and the country needs loyalty to leaders.
Belief that the will of people can be stopped, that it's unacceptable for a strong leader to be above the law and preference for direct democracy and referendums.
Lower interest in politics and lower confidence in political participation.
Strong opposition to reducing income inequality.
Disagreement with the freedom of gays and lesbians.
Belief that climate change is not happening or is not human caused.
Varied opinions on the origins of COVID-19.

Table 6.10: Characteristics of participants that increase nonresponse probability.

Demographic Characteristics of nonresponders
More likely to be female.
More likely to live in rural areas.
Less likely to belong to a discriminated group.
Less likely to live in multi-person households.
More likely to be non-citizens of the country.
More likely to have completed high school or university.
More likely to have low income.

Table 6.11: Demographic characteristics of nonresponders.



This profiling indicates that nonresponse to the voting question is associated with specific trust levels, satisfaction, political engagement, socio-economic factors, and demographic characteristics, supporting a MAR (Missing at Random) pattern.

6.7. Implementation of Clustering Method

6.7.1. Implementation

In order to delve into nonresponders characteristics and identify special patterns, we decided to perform clustering. As previously mentioned in Chapter 3, clustering was implemented using the [BayesBinMix](#) package (Papastamoulis & Rattray, 2017). The independent variables we used for the implementation were a subset of 15 variables from the 38 that we kept in *Model 4*. We kept those 15 according to their p-value, their statistical significance in the model.

The variables included in the clustering implementation:

- *stfgov*: How satisfied with the national government
- *votedir*: Citizens have the final say on political issues by voting directly in referendums
- *fairelcc*: In country national elections are free and fair – 0 does not apply at all :10 applies completely
- *dfprtalc*: In country different political parties offer clear alternatives to one another
- *rghmgprc*: In country the rights of minority groups are protected – 0 does not apply at all.
- *gtpelcc*: In country governing parties are punished in elections when they have done a bad job.
- *gvtzpv*: In country the government protects all citizens against poverty.
- *wpestopc*: In country the will of the people cannot be stopped.
- *accalaw*: Acceptable for country to have a strong leader above the law.
- *panfolru*: More important to follow government rules or to make own decisions when fighting a pandemic.
- *freehms*: Gays and lesbians free to live life as they wish.
- *loylead*: Country needs most loyalty towards its leaders
- *ccnthum*: Climate change caused by natural processes, human activity, or both



- *gvconc19*: COVID-19 is result of deliberate and concealed efforts of some government or organization.
- *dscrgrp*: Member of a group discriminated against in this country

Since the BayesBinMix package performs clustering for binary data with missing values, we used the original data set (before the imputation) and we created a data set containing dummy variables (0-1) using the [fastDummies](#) (Kaplan & Schlegel, 2023) package in R. We performed clustering for a sample of our observations. Our sample consists of all nonresponders (524 observations) and a random sample of 550 responders (we set seed “123” for reproducibility reasons). The final matrix was 1074 x 132. We used the main function of the [R package](#) called *coupledMetropolis*. This function takes as input a binary data array, in our case containing missing values, and runs the allocation sampler for a series of heated chains which run in parallel while swaps between pairs of chains are proposed (Papastamoulis & Rattray, 2017). Before implementing the *coupledMetropolis* we set seed “123” for reproducibility reasons. The arguments of the function were specified as follows:

- $K_{max} = 10$ the maximum number of clusters
- $nChains = 8$ parallel heated chains.
- $heats = (1.00, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70, 0.65)$ the temperature vector was chosen in order for the swap acceptance to be $>20\%$.
- $ClusterPrior = poisson$. We chose the truncated Poisson over the uniform because it is suggested by Papastamoulis as Rattray (2017). Nevertheless, we performed clustering using also the uniform as cluster prior, the results were the same.
- $alpha = 1$, first shape parameter of the Beta distribution.
- $beta = 1$, second shape parameter of the Beta distribution. We tried both uniform distribution and Jeffrey’s prior ($\alpha = \beta = 0.5$) as prior distribution for the Bernoulli parameters as suggested (Papastamoulis & Rattray, 2017).
- $m = 8800$
- $burn = 800$

The swap acceptance rate was 24%. The most probable value of K is 2 as shown in the table below. Most probable model: $K = 2$ with $P(K = 2 | data) = 0.996$.



Estimated posterior distribution of the number of clusters	
2	3
0.996	0.004

Table 6.12: Estimated posterior distribution of the number of clusters

Estimated number of observations per cluster conditionally on $K=2$ (3 label switching algorithms).			
	STEPHENS	ECR	ECR.ITERATIVE.1
1	735	735	735
2	339	339	339

Table 6.13: Estimated number of observations per cluster conditionally on $K = 2$

In Table 6.13 we observe how observations are distributed in the two clusters. More than the half of the observations belong to cluster 1.

6.7.2. Interpreting Clustering Results

In this section we will attempt to interpret the clustering of nonresponders. Our goal is to identify patterns which will enlighten the characteristics of nonresponders. Given that the vote choice of responders is available, we created a two-way frequency table to represent how the different votes are distributed in the clusters (Table 6.14).

Party voted	Cluster 1	Cluster 2
New Democracy (ND)	107	153
SYRIZA	116	29
KINIMA ALLAGIS (KIN.ALL, f. PASOK)	40	15
KKE (Communist Party of Greece)	26	5
Elliniki Lisi	6	7
Mera25 (Diem25)	12	1
Golden Dawn	14	1
Plefsi Eleftherias	3	2
Enosi Kentroon	3	2
Other	8	0

Table 6.14: Two - way frequency table "party voted for" vs "cluster".



In Table 6.14 we observe that cluster 1 has fewer voters of New Democracy compared to cluster 2. Cluster 1 has obviously more voters of SYRIZA, KIN. ALL, KKE, MERA 25 and Golden Dawn compared to cluster 2. The voters of Elliniki Lisi, Eleftherias Pleyisi Eleftherias and the Enosi Kentroon are divided between the two clusters. Regarding the nonresponders we can see how they are distributed in the two clusters. Most of them (more than half) belong in cluster 1.

Cluster 1	Cluster 2
401	123

Table 6.15: Distribution of non - responders in the two clusters.

We calculated the covariance and correlation matrix of the model we obtained after performing the clustering to examine the patterns in the model. To do so we used the formula below given by Bishop (2006, p. 445)

$$cov[x] = \sum_{\kappa=1}^K \pi_{\kappa} \{ \Sigma_{\kappa} + \mu_{\kappa} \mu_{\kappa}^{\tau} \} - E(X)E(X)^{\tau}$$

Where:

$$E(x) = \sum_{\kappa=1}^K \pi_{\kappa} \mu_{\kappa}$$

$$\Sigma_{\kappa} = diag\{\mu_{\kappa i}(1 - \mu_{\kappa i})\}$$

According to the annotation we use $K = 2$, $\mu_1 = \theta_1$ and $\mu_2 = \theta_2$.

The values in the correlation matrix are equal to zero or almost equal to zero meaning that the selected variables for clustering nonresponders and a sample of responders do not exhibit strong linear relationships within the clusters. This can imply the presence of more complex interactions among the variables that define the clusters. The identified clusters are meaningful in differentiating groups despite the lack of linear correlations among the variables.

In Figure 6.9 we see the probability of choosing the j_{th} level of each categorical independent variable in each of the two clusters. The probabilities are presented in descending order with respect to the absolute difference between the two clusters. With the straight lines we represent the difference. We will use this figure to extract assumptions concerning the differences in participants' characteristics between the two clusters.



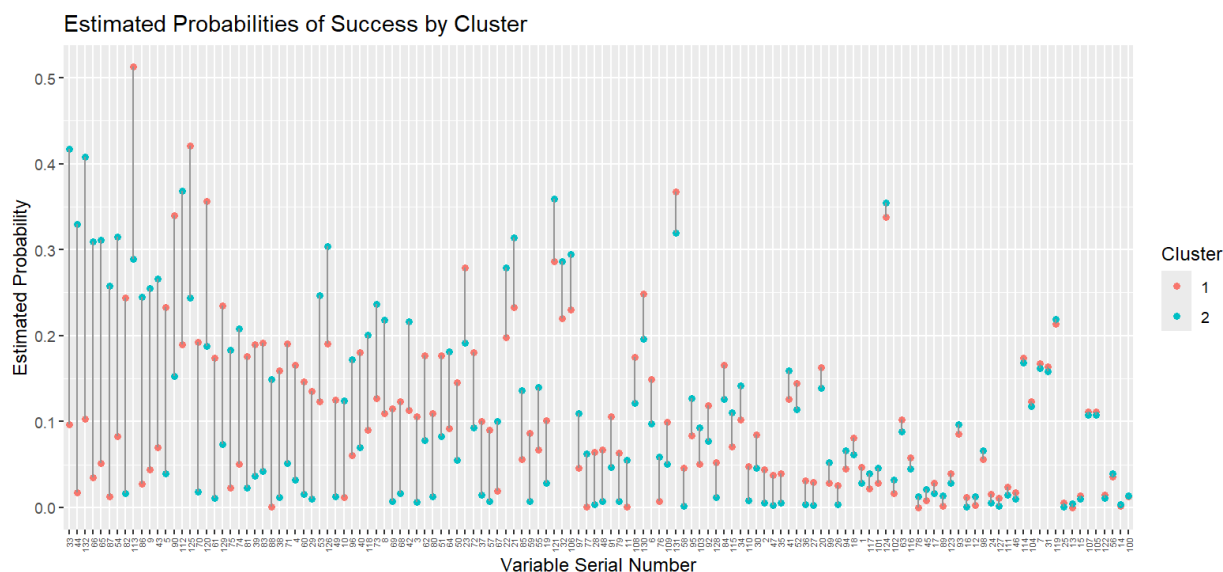


Figure 6.9: The probability of choosing the j level of each categorical variable in each of the two clusters.

We annotated the 20 largest differences to identify interesting patterns for profiling nonresponders.

Cluster 1
8.in Greece the will of the people can be stopped. ($\theta_{82,1}=0.24$)
9.someone agrees with the view that gays and lesbians have the right to live life as they wish. ($\theta_{113,1}=0.51$)
13. not quite satisfied with the government (ND). ($\theta_{5,1}=0.23$)
14. does not agree that it is acceptable for a country to have a leader above the law. ($\theta_{90,1}=0.34$)
16.climate change is caused by human activity. ($\theta_{125,1}=0.42$)
17.in Greece the government does not protect all citizens against poverty. ($\theta_{70,1}=0.19$)
18.disagrees that the country needs most loyalty towards its leaders. ($\theta_{120,1}=0.36$)
19.in country governing parties are not punished if they have done a bad job. ($\theta_{61,1}=0.17$)
18.agreeing with the view that COVID-19 was the result of deliberate and concealed efforts of governments or organizations. ($\theta_{129,1}=0.23$)

Table 6.16: Characteristic of participants in cluster 1 according to the difference in the probabilities of choosing a level j from a covariate between the two clusters.



Cluster 2
1.it is true that in Greece national elections are free and fair. ($\theta_{33,2}=0.41$)
2.in Greece different political parties offer different alternatives. ($\theta_{44,2}=0.33$)
3.strongly disagreeing that COVID-19 was the result of deliberate and concealed efforts of governments or organizations. ($\theta_{132,2}=0.41$)
4.in country governing parties are punished when they have done a bad job (level 9 at the scale from 0 – 10, 0:completely disagree and 10: completely agree). ($\theta_{66,2}=0.31$)
5.in country governing parties are punished when they have done a bad job (level 8 at the scale from 0 – 10). ($\theta_{65,2}=0.31$)
6.in Greece the will of the people cannot be stopped. ($\theta_{87,2}=0.31$)
7.the rights of minority groups are protected in Greece. ($\theta_{54,2}=0.26$)
10.in country governing parties are punished when they have done a bad job (level 7 at the scale from 0 – 10). ($\theta_{86,2}=0.25$)
11. someone is relatively satisfied with the government (ND). ($\theta_{9,2}=0.26$)
12.in Greece different political parties offer different alternatives. ($\theta_{43,2}=0.27$)
15. strongly agrees that gays and lesbians have the right to live life as they wish ($\theta_{112,2}=0.37$)

Table 6.17: Characteristic of participants in cluster 2 according to the difference in the probabilities of choosing a level j from a covariate between the two clusters.

As we observed in Table 6.15, the majority of nonresponders belong to cluster 1. We presented in Table 6.16 and Table 6.17 the differences in the probabilities of choosing a level from a covariate between cluster 1 and cluster 2 and we can make assumptions now. An interesting characteristic which separates those belonging to cluster 1 with those in cluster 2 is the satisfaction with the government. Those belonging to cluster 1 tend to be less satisfied with the government compared to those belonging to cluster 2. Moreover, people in cluster 2 believe that in Greece when a governing party has done a bad job is punished while in cluster 1, they believe that this is not true. In general, we observe that people belonging in cluster 1 express dissatisfaction towards the government this can explain Table 6.14 where we observe more votes for the government party to belong in cluster 2. In addition, cluster 2 is characterized by opinions in favor of the LGBTQI+ community, against authoritarianism.



6.8. Conclusion

In the chapter above we explained how we implemented our methods. We firstly imputed the missing values in the independent variables, then we performed variable selection using LASSO, group-Lasso and stepwise with AIC and we identified the significant variables. We finally presented the characteristics of nonresponders and we explored the MAR assumption.





CHAPTER 7

Conclusion

7.1. Introduction

Concluding our thesis, we will review our objectives, our approach, and the methods we used. We will highlight the successes, the challenges and the difficulties we faced when conducting our analysis. Finally, we will extract a conclusion about the voting choice and characteristics of nonresponders.

In our analysis, our focus centers on nonresponse related to a question about the voting behavior of the interviewees. To be more precise, we attempted to construct the profile of those individuals who refused to disclose their voting choices in the last national election and identify their characteristics. Our aim was to contribute to the bibliography and research concerning item nonresponse in electoral studies, a topic that has posed challenges to scientists for many years.

Year by year, the number of individuals who choose not to respond to surveys and vote-related questions increases. Scientists study the phenomenon from different perspectives. There is a discussion around response rates and their relation to nonresponse bias. As we explained above nonresponse bias could occur when we have differences in the characteristics of responders and nonresponders. This is why we care about the profile of nonresponders to weight for them properly. Moreover, the discussion includes the topic of increasing response rates. This issue lies in different aspects of a survey. The questionnaire's size, the type of questions, the interview's characteristics and the interviewer's attitude, are some aspects studied in the last years from scientists. The goal is to find the optimal ways to create the questionnaire and at the same time educate the interviewers properly to get the maximum results.

In our thesis, we discussed the first problem of nonresponse bias, and especially the problem of finding out which are the characteristics of people who choose not to reveal what they voted for. We start by assuming that our data are Missing At Random and as a consequence we have enough information to predict the missing answers.

7.2. Overcoming Research Difficulties

In Chapter 2 we discussed methods which were developed to describe the profile or allocate undecided voters or to predict their voting choices in pre-election surveys. However, our problem is slightly different since we aim to cluster or estimate



the vote choice in a post-election survey. In our study, participants have already voted; therefore, they are not considered as undecided voters. Their choice not to respond is intentional and for the majority of nonresponders, we could estimate their voting choice. We use the term ‘majority’ because, according to the literature, some vote overreporting is expected. Individuals who falsely claim to have voted may choose to refuse answering the question about their voting choice. Moreover, we need to account for social desirability bias, we expect some individuals to withhold their voting choice in 2019’s national election. For example, voters of Golden Dawn, a party that was convicted as a criminal organization in 2020, may prefer to hide what they voted for.

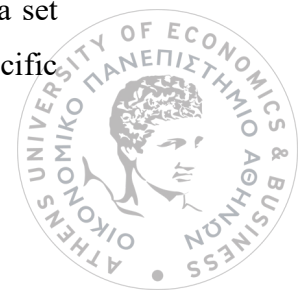
In our study we face another difficulty: the missing values across our data set. All individuals chose not to answer one or more questions which could be useful to estimate the nonresponses in vote choice question. Item nonresponse was a phenomenon which appeared in most of the participants, not only the ones who chose not to answer what they voted for. Listwise deletion of missing values in independent variables would mean loss of useful information.

Missing values in our independent variables was a challenge and a risk. In general, large-scale surveys like the ESS, due to the volume of the questionnaire is more probable to have multiple missing values. At the same time, the length of the ESS had as consequence large data dimensions which had to be dealt smartly and carefully. We can assume, that maybe a pre-election survey would be easier to handle due to the smaller size and the content. We think that it would be easier to identify undecided voters’ profile and that missing values in independent variables would be rare.

Another difficulty and challenge we confronted was the type of our data in combination with the dimensions of the data set. Most of our variables were either in Likert scale or just categorical ones. We had to decide which to treat as integer numeric variables, which to treat as ordinal factors and which as factors. Then we had to decide if ordinal variables would be used in the analysis as continuous or not. Data dimensions and multilevel categorical data lead to difficulties in calculations, model building, model selection and clustering. This led us to the next topic of discussion, data cleaning.

7.3. An Overview of Data Preparation Process

An important aspect of our study was the data preparation process. As we already explained we had to deal with large data dimensions. Moreover, the data set included auxiliary variables, test questions or questions that were asked under specific



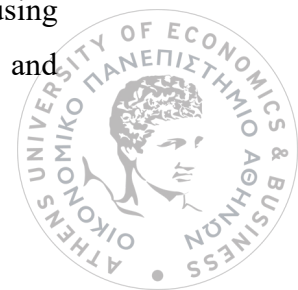
circumstances. We began data cleaning by discarding auxiliary and test questions as well as questions that were not asked to the whole sample. The next steps were more complicated and demanded synthetic thinking and decision-making. First of all, we created new scores by merging variables into one when possible. Secondly, we transformed variables to reduce levels' dimensions and lastly, we omitted variables which we thought did not offer valuable information for our analysis. The next step was to decide the mode of our variables. The only challenge we faced concerned the ordinal variables, to be more precise we had to decide if to treat them as factors or continuous. After presenting both arguments, we decided to follow the two approaches and adopt the optimal one.

Another approach we could follow to reduce data dimensions is Factor Analysis. It is a commonly used technique in social survey data as it combines related questions in one factor to provide more robust measures. This practice could be implemented in future analysis of the ESS Round 10 data.

7.4. Evaluation of the Methods

One of the main goals of our analysis was to investigate if our data are MAR with respect to the question about voting choice. The MAR assumption would mean that the missing values in the *prvtmgr* variable are related to variables in the dataset and that we could use them in order to estimate the missing values. To achieve this first goal, we decided to fit a logistic regression model, with response a Y:0-1 response and nonresponse respectively. We fitted the model and performed variable selection to find which covariates are significant and as a consequence are related to nonresponse. Besides creating a logistic model which could support a MAR pattern, this method leads us to the conclusion that the differences between responders and nonresponders are significant. Additionally, using the model and implementing the coefficients we created the profile of nonresponders.

To begin with, we had to solve the problem of missing values in the independent variables. We chose to implement Multiple Imputation (MI) with the AMELIA (King et al., 2011) package in R, a technique suitable for political surveys' data. As a plus the AMELIA can perform MI to factor variables as well as to numeric ones and it was suitable for our data set. Secondly, we used the imputed data set to fit the logistic model. Given the dimensions of our data set, we chose to perform variable selection using LASSO (Tibshirani, 1996) screening techniques to avoid multicollinearity and



overfitting. Since our data contained numerous factor variables group-lasso (Yuan & Lin, 2005) emerged as a logical consequence. It is a generalization of LASSO for groupwise variable selection. This method is suitable for categorical data since it encourages entire groups of coefficients to be zero and it selects or excludes entire groups of variables. We implemented both LASSO and group-lasso and compared the methods. It is important to note that LASSO retained slightly less covariates than group-lasso and it was faster. Nevertheless, the models fitted after group-lasso performed better and exhibited smaller AIC. Thirdly, we decided to perform variable selection, for the three models, using the stepwise method and the AIC criterion. This method was time consuming, because of the many factor variables which increased models' dimensions. We ended up with four models with 43,41, 36 and 38 covariates respectively.

The next step was to evaluate the models. We chose to perform 5-fold cross validation and create ROC curves for each fold to evaluate models' performance. The results were alarming, all five AUC (area under the curve) for the three models, were over 0.5 but below 0.8 (greater than 0.8 for the performance to be acceptable). None of our models performed above this threshold. The AUC values for our models indicated fair diagnostic accuracy but left room for improvement. After comparing the AUC values for all the models, we decided that *Model 4*, the model which assumes ordinal covariates as factors, had the best performance. We must underline, that this model also demonstrates the most favorable AIC value amongst the three models. To conclude, although none of the models reach the threshold for excellent discrimination, the chosen model provides a fair and consistent approach for predicting nonresponse, which can be valuable in practical applications where perfect accuracy is not always achievable.

The last method we used was clustering. We performed clustering to better understand nonresponders' characteristics and explore the existence of special patterns. The package we used to implement clustering was the BayesBinMix (Papastamoulis & Rattray, 2017) package in R. We chose this package because it is suitable for data sets with missing values. Moreover, the BayesBinMix performs model-based clustering for multivariate binary data, which was also appropriate for the type of our covariates (factors). We created a new data set turning our covariates to dummy variables. The BayesBinMix utilizes MCMC sampling to jointly estimate the number of cluster and the model parameters. Specifically, Papastamoulis and Rattray (2017), use the



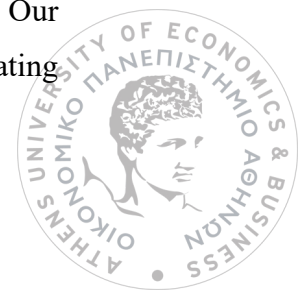
Metropolis – coupled MCMC (MC3) sampler. MC3 which also utilizes parallel heated chains. Since the MCMC algorithm is computationally challenging and time consuming we had to use a subset of our data set. We used 15 variables from the 39 that we kept in *Model 4*. We kept those 15 according to their p-value, their statistical significance in the model. We also kept a sub-sample of our observations. To elaborate this, the sub-sample consisted of all non - responders (524 observations) and a random sample of 550 responders. The final matrix was 1074 x 132. This clustering resulted in 2 clusters, but it is important to keep in mind that this result cannot be generalized in the whole sample and maybe the results would be different if we used all the observations and all the covariates.

7.5. Was it Worth the Wait? Analyzing the Results and Validating the Assumptions.

The methods we explained and evaluated above lead to some results. The goal was to prove that the missing values in the *party voted question* were dependent to the rest of the variables and to the answers the participants gave to the rest of the questions (MAR assumption). In addition, we aimed in creating the profile of non – responders, to identify those characteristics that discriminate them from the responders.

7.5.1. Step 1: Logistic Model Results

The logistic model showed that there are variables from our data set related with the nonresponse. A MAR pattern could be supported by fitting the logistic regression model and finding the significant variables which influenced the nonresponse (Assumption 4). Additionally, the logistic model revealed some of the characteristics of individuals who did not respond to the voting choice question in the last national elections and their differences from those that tend to respond (Assumption 4). Consequently, the model indicates that nonresponse to the voting question is associated with specific trust levels, satisfaction, political engagement, socio-economic factors, and demographic characteristics (Assumption 1). To be more precise, as the literature suggests (Brehm, 1987; Alexander, 2018; Matsuo et al., 2018), our model proved that nonresponders are in fact, are more likely to be less interested and less engaged in politics. In addition, income was a significant covariate for nonresponse, since the wealthier the individual, the more probable the nonresponse. Another factor that influences nonresponse according to our model and the literature was the area of residence, we discovered that nonresponders are more likely to live in a village. Our research proved that women and foreigners are more prone to nonresponse validating



Alexander's (2018) findings. As Matsuo et al. (2018) suggest, our model showed that nonresponders are less satisfied with the functioning of democracy. In contrast with the literature, we did not find that age was a significant covariate for nonresponse.

The ROC curves we computed, to evaluate the logistic model's predictive performance, proved that our model is meaningful, since it predicts more accurate than a simple random guess. Given that, this model is a useful tool to predict nonresponse and identify the characteristics of nonresponders.

Another interesting conclusion is that the set of characteristics of nonresponders highlights different ideologies, social and political views, which makes it almost impossible to draw a specific conclusion about their political and ideological identity. This can be explained if we consider that those who do not answer this question do not constitute a homogeneous group with common characteristics. On the contrary, like the whole of society and as part of society, they differ from each other.

7.5.2. An in-depth exploration of this special group

As already mentioned, the next step was to discover differences amongst the nonresponders. This would prove that even this group is not homogeneous and nonresponders could express different views in several social and political issues. To prove that we utilized a clustering technique, BayesBinMix. This technique resulted in 2 clusters in the subsample we used for implementation. Nonresponders were unevenly split in both clusters (401 in cluster 1, 123 in cluster 2) (Assumption 5).

We started by studying the distribution of the sample of responders in the two clusters, and in particular how they are distributed by their vote. The first cluster has fewer voters of ND compared to the second and more of SYRIZA, PASOK, KKE, MERA25, Golden Dawn compared to the second cluster. Could we assume an extension of this distribution of voters to those who do not respond?

Exploring the characteristics that differentiate those belonging to the first cluster from those belonging to the second cluster, we analyzed the top 20 differences in the probabilities of choosing a level from a covariate between the two clusters. The first cluster is characterized by critical attitudes towards government and governability in general. At the same time, it is governed by environmental concerns and pro-LGBTQ social sensitivities. Those belonging in the second cluster, have more trust in the democratic process and the effectiveness of political accountability in Greece. They



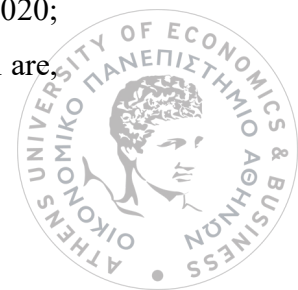
have a higher level of satisfaction with the current government, while firmly rejecting conspiracy theories related to COVID-19.

The above analysis can give us some insights into the characteristics of nonresponders as well as confirm the hypothesis that they are a heterogeneous group. Indeed, the separation into two clusters highlights the heterogeneity of nonresponders that is consistent with the heterogeneity of the population. A future cluster analysis using the whole sample may lead to more detailed conclusions.

7.5.3. Realistic Scenario or Elusive Goal: Predicting the Vote of Nonresponders

The aim of many social and political scientists is of course to mitigate the issue of nonresponse. We have presented the risks and challenges of different kinds of nonresponse. Here we are concerned with item nonresponse and, as we explained, attempts are constantly being made to identify the characteristics of those who do not respond as well as to predict their responses. A typical example is voting estimation, something we see very often in polls especially in recent years when the percentage of "undecided" is increasing. We presented several methods on the prediction or allocation of undecideds and those who conceal what they voted (Fenwick et al. 1982; Visser et al., 2000; King et al., 2001; Nandram and Choi, 2008; Liu et al., 2021). The existence of missing values in our independent variables made it nearly impossible to apply the above methods. At the same time, we were faced with another challenge in this research. The elections, to which the question that concerns us refers, were held in 2019 while the data collection took place three years later in 2022. During these years, much has changed in the Greek political scene. The views of many survey participants around various issues are likely to have changed and predicting the 2019 vote based on responses and opinions in 2022 may not be meaningful. Nevertheless, the logistic model we created to estimate nonresponse in the voting question, performs better than the random prediction as we proved by computing the ROC curve. The use of a logistic model to predict nonresponse and identify the characteristics of nonresponders is meaningful.

At the same time, we should comment that knowledge of electoral behavior and sociology would help us to better understand our results and draw conclusions about the electoral and party identity of the participants, using also the characteristics of political parties and their voters as derived from various studies (Mylonas, 2020; Tsatsanis et al., 2020; Elias et al., 2022). Interdisciplinary study and collaboration are,



after all, characteristics of statistics, a discipline that is at the intersection of different fields.

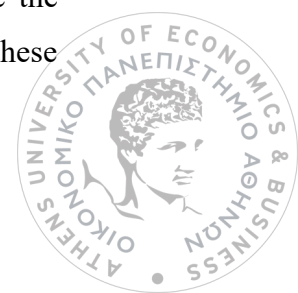
7.6. Conclusion of the Conclusion

In conclusion, our study aimed to profile nonresponders concerning their voting behavior in the 2019 national election in Greece. Through various methodological approaches, we highlighted the complex nature of nonresponse bias, and we identified key characteristics that distinguish nonresponders from responders. Our analysis emphasized the importance of understanding nonresponders to reduce and deal with nonresponse bias effectively.

We faced several challenges and difficulties, like dealing with missing data and the large dimensionality of our dataset. We employed techniques like Multiple Imputation and variable selection methods, and we managed to derive meaningful insights despite these obstacles.

Our logistic regression model identified significant covariates that influence nonresponse, such as political engagement and interest, socio-economic factors, and demographic characteristics. These findings partially align with the existing literature and highlight the multifaceted reasons behind individuals' reluctance to disclose their voting choices. Additionally, the clustering analysis provided an in-depth exploration of nonresponders and showed distinct subgroups with varying political attitudes, levels of trust in the democratic process and satisfaction with the government. The clustering analysis revealed that nonresponders are a heterogeneous group, that reflects the broader diversity of the population. This finding underscores the need for tailored strategies when addressing item nonresponse in electoral studies. Moreover, it highlights the difficulty of resolving the common problem of item nonresponse and subsequently of nonresponse bias. The fact that nonresponders do not form a homogeneous group makes it very difficult to mitigate the problem, as the puzzle of how to weight for this group remains unsolved.

Looking forward, future research could benefit from applying these methods and at the same time exploring other approaches like Factor Analysis. In addition, cluster analysis, which can shed light to the distinct groups and characteristics of nonresponders, should be applied to the whole data set using all covariates and observations. It is important to note that *BayesBinMix* can be used to estimate the missing values in the independent variables and future research can compare these



results with Multiple Imputation. Interdisciplinary collaboration, particularly involving experts in electoral behavior, political science and sociology, could enrich the analysis and lead to more comprehensive conclusions.

Overall, our study contributes to the ongoing discourse on nonresponse in electoral and social surveys and offers valuable insights and methodological advancements to better understand and address this persistent challenge in survey research.





References

- Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Applied Statistics* 57 (3), pp. 273-291.
- Alexander, E. C. (2018). Don't Know or Won't Say? Exploring How Colorblind Norms Shape Item Nonresponse in Social Surveys. *Sociology of Race and Ethnicity*, 4(3), pp. 417-433.
- Anseel, F., Lievens, F., Schollaert, E., & Choragwicka, B. (2010). Response Rates in Organizational Science, 1995–2008: A Meta-analytic Review and Guidelines for Survey Researchers. *Journal of Business and Psychology*, 25(3), pp. 335-349. doi:10.1007/s10869-010-9157-6
- Berinsky, A. J. (2002). Silent Voices: Social Welfare Policy Opinions and Political Equality in America. *American Journal of Political Science*, 46(2). doi:10.2307/3088376
- Bethlehem, J. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. In R. Groves, D. Dillman, J. Elinge, R. Little, R. Groves, D. Dillman, J. Elinge, & R. Little (Eds.), *Survey Nonresponse* (pp. 275-288). New York: Wiley.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.
- Brehm, J. (1987). Who's Missing? An Analysis of Nonresponse and Undercoverage in the 1986 National Election Studies Post-election Survey. *Working Paper, No 10*.
- Donders, A. R., Heijden, G. J., Stijnen, T., & Moonsf, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59, pp. 1087-1091.
- Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, 37(2), pp. 90-93.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32, pp. 407-451. Retrieved from <http://www.jstor.org/stable/3448465>



- Elias, D., Costas, E., Alexia, K., George, K., André, K., Yordan, K., . . . Eftichia, T. (2022). The Greek Political Landscape 2019 - 2021. (E. Teperoglou, Ed.) *Friedrich-Ebert-Stiftung Athens Office*.
- Enders, C. K. (2022). *Applied Missing Data Analysis*. New York: The Guilford Press.
- European Social Survey*. (n.d.). Retrieved from <https://www.europeansocialsurvey.org/about-ess>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp. 861-874. doi:<https://doi.org/10.1016/j.patrec.2005.10.010>.
- Fenwick, I., Wiseman, F., Becker, J. F., & Heiman, J. R. (1982). Classifying Undecided Voters in Pre-Election Polls. *The Public Opinion Quarterly*, Vol. 46, No. 3, pp. 383-391.
- Fulton, B. R. (2018). Organizations and Survey Research: Implementing Response Enhancing Strategies and Conducting Nonresponse Analyses. *Sociological Methods & Research*, 47(2), pp. 240-276. doi:<https://doi.org/10.1177/0049124115626169>
- Graham, J. W. (2012). *Missing Data: Analysis and Design*. New York: Springer.
- Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5), pp. 646-675. doi: <https://doi.org/10.1093/poq/nfl033>
- Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *The Public Opinion Quarterly*, 72(2), pp. 167-189. Retrieved from <https://www.jstor.org/stable/25167621>
- Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in Producing Nonresponse Bias. *The Public Opinion Quarterly*, 70, pp. 720-736. Retrieved from <https://www.jstor.org/stable/4124223>
- Hastie, T., Qian, J., & Tay, K. (2023). *Lasso and Elastic - Net Regularized Generalized Linear Models*. Retrieved from An Introduction to glmnet: <https://glmnet.stanford.edu/articles/glmnet.html>



- Heer, W. d. (1999). International Response Trends: Results of an International Survey. *Journal of Official Statistics, Vol. 15, No. 2*, pp. 129-142.
- Hoek, J., & Gendal, P. (1997). Factors Affecting Political Poll Accuracy: An Analysis of Undecided Respondents. *Marketing Bulletin*, pp. 1-14.
- Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science, 20*, pp. 50-67. doi:10.1214/088342305000000016
- Kaplan, J., & Schlegel, B. (2023). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*. Retrieved from <https://www.rdocumentation.org/packages/fastDummies/versions/1.7.3>
- King, G., Honaker, J., & Blackwell, M. (2011). Amelia II: A Program for Missing Data, Volume 45, Issue 7. *Journal of Statistical Software*.
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*.
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Qual Quant 47*, pp. 2025-2047.
- Lall, R. (2016). How Multiple Imputation Makes a Difference. *Political Analysis*., 24(4). doi:10.1093/pan/mpw020.
- Liu, Y., Ye, C., J. S., Jiang, Y., & Wang, H. (2021). Modeling undecided voters to forecast elections: From bandwagon behavior and the spiral of silence perspective. *International Journal of Forecasting 37*, pp. 461-483.
- Matsuo, H., Billiet, J., Loosveldt, G., & Berglund, F. (2018). Measurement and adjustment of nonresponse bias based on nonresponse surveys: the case of Belgium and Norway in the European Social Survey Round 3. *Survey Research Methods, Vol. 4, No. 3*, pp. 165-178.
- Mylonas, H. (2020). Greece: Political Developments and Data in 2019. *European Journal of Political Research Political Data Yearbook, 59(1)*, pp. 1-14. doi:10.1111/2047-8852.12299



- Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1), pp. 25-36. doi:<https://doi.org/10.4097/kja.21209>
- Nandram, B., & Choi, J. W. (2008). A Bayesian allocation of undecided voters. *Survey Methodology*, Vol. 34, No.1 , pp. 37-49.
- Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication*, 24(2), pp. 43-51. doi:<https://doi.org/10.1111/j.1460-2466.1974.tb00367.x>
- Papastamoulis, P., & Rattray, M. (2017). BayesBinMiz: an R Package for Model Based Clustering of Multivariate Binary Data. *The R Journal*, 9(1), pp. 403-420. Retrieved from <https://journal.r-project.org/archive/2017/RJ-2017-022/index.html>
- Peytchev, A. (2013). Consequences of Survey Nonresponse. *The Annals of the American Academy of Political and Social Science*, 645, pp. 88-111. Retrieved from <https://www.jstor.org/stable/23479083>
- Piekut, A. (2019). Survey nonresponse in attitudes towards immigration in Europe. *Journal of Ethnic and Migration Studies*.
- Robitzsch, A. (2020). Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods. *Frontiers in Education* , 5. doi:10.3389/feduc.2020.589965
- Rogelberg, S. G., & Stanton, J. M. (2007). Introduction: Understanding and Dealing With Organizational Survey Nonresponse. *Organizational Research Methods*, 10(2), pp. 195-209. doi:10.1177/1094428106294693
- Silver, B. D., Anderson, B. A., & Abramson, P. R. (1986). Who Overreports Voting? . *American Political Science Review*, 80(2), pp. 613-624.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58, pp. 267-288. Retrieved from <https://www.jstor.org/stable/2346178>



- Tsatsanis, E., Teperoglou, E., & Seriatos, A. (2020). Two-partyism Reloaded: Polarisation, Negative Partisanship, and the Return of the Left - right Divide in the Greek Elections of 2019. *South European Society and Politics*. doi:10.1080/13608746.2020.1855798
- Turk, A., Heneghan, C., & Nunan, D. (2019). *Catalogue of Bias*. Retrieved from Nonresponse bias: <https://catalogofbias.org/biases/nonresponse-bias/>
- Valentino, N. A., Hill, W. W., & King, J. L. (2017). Polling and Prediction in the 2016 Presidential Election. *Computer*, vol. 50, no. 05, pp. 110-115.
- Visser, P. S., Krosnick, J. A., Marquette, J. F., & Curtin, M. F. (2000). Improving Election Forecasting: Allocation of Undecided Respondents, Identification of Likely Voters, and Response Order Effects. In P. Lavrakas, & M. Traugott, *Election poll, the news media, and democracy*. New York: NY: Chatham House.
- Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalized learning problems. *Stat Comput*(25), pp. 1129-1141. doi:<https://doi.org/10.1007/s11222-014-9498-5>
- Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, 93(1). doi:10.1002/ecm.1557
- Yuan, M., & Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1), pp. 49-67. doi:<https://doi.org/10.1111/j.1467-9868.2005.00532.x>
- Βούλγαρης, Γ., & Νικολακόπουλος, Η. (2014). *2012: Ο Διπλός Εκλογικός Σεισμός*. Αθήνα: ΘΕΜΕΛΙΟ.





Appendix

List of tables

edlevegr	edrec
1: Merikes taxeis Dimotikou	Lower secondary and less
2: Apolytirio Dimotikou	Lower secondary and less
3: Apolytirio Gymnasiou	Lower secondary and less
4: Pistopoiitiko Epangelmatikis Katartisis epipedou 1	Lower secondary and less
5: Apolytirio Genikou Lykeiou	Advanced Secondary
6: Ptychio Epangelmatikis Ekpedefsis epipedou 3	Advanced Secondary
7: Apolytirio Epaggelmatikou Lykeiou kai Ptychio Epipedou 3	Advanced Secondary
8: Diploma epangelmatikis katartisis epipedou metadefterovathmias epangelmatikis katartisis	Advanced vocational
9: Ptychio (Sxoles Anoteris Epaggelmatikis Ekpaidefsis)	Tertiary
10: Ptychio (ATEI)	Tertiary
11: Ptychio (AEI)	Tertiary
12: Metaptychiako diploma idikefsis (ATEI)	Tertiary
13: Metaptychiako diploma idikefsis (AEI)	Tertiary
14: Metaptychiako Diploma Idikefsis from Polytechnical School, Agricultural schools etc	Tertiary
15: Didaktoriko Diploma	Tertiary
5555: other	other
NA	NA

Table 4.1: Recoding of the edrec variable.

Basic descriptive statistics					
<i>Variable</i>	<i>Mean</i>	<i>SD</i>	<i>Median</i>	<i>Range</i>	<i>Skewness</i>
nwspol	107.87	125.52	60	1205 (0-1205)	4.45
trstprl	4.38	2.28	5	10 (0-10)	-0.09
trstlgl	6.54	2.22	7	10 (0-10)	-0.74
trstplc	6.84	2.11	7	10 (0-10)	-0.79
trstplt	3.70	2.18	4	10 (0-10)	0.30
trstprt	3.59	2.13	4	10 (0-10)	0.29
trstep	4.60	2.17	5	10 (0-10)	-0.12
trstun	4.64	2.25	5	10 (0-10)	-0.12
trstsci	7.37	1.90	8	10 (0-10)	-1.09
stflife	6.38	1.71	7	10 (0-10)	-0.62
stfeco	3.87	2.13	4	10 (0-10)	0.18
stfdem	5.09	2.23	5	10 (0-10)	-0.29
stfedu	4.56	2.08	5	10 (0-10)	0.03
stfhlth	4.61	2.33	5	10 (0-10)	-0.04
eufff	5.04	2.18	5	10 (0-10)	-0.20
imwbcnt	4.26	2.15	4	10 (0-10)	0.07



happy	6.61	1.52	7	10 (0-10)	-0.78
atchctr	8.85	1.41	9	10 (0-10)	-1.58
atcherp	5.46	2.27	6	10 (0-10)	-0.38
rlgdgr	6.36	2.23	7	10 (0-10)	-0.74
fairelc	9.02	1.35	10	10 (0-10)	-1.92
dfprtal	8.44	1.47	9	9 (1-10)	-1.26
medcrgv	8.61	1.52	9	10 (0-10)	-1.49
rghmgrpr	8.24	1.69	9	10 (0-10)	-1.38
cttresa	8.79	1.44	9	10 (0-10)	-1.50
gptpelc	8.51	1.44	9	10 (0-10)	-1.24
gvctzpv	8.71	1.49	9	10 (0-10)	-1.53
grdfinc	8.39	1.65	9	10 (0-10)	-1.38
viapol	8.02	1.69	8	10 (0-10)	-0.87
wpestop	8.61	1.41	9	10 (0-10)	-1.23
keydec	7.99	1.72	8	10 (0-10)	-0.90
medcrgvc	5.07	2.64	5	10 (0-10)	-0.12
votedirc	3.80	2.43	4	10 (0-10)	0.37
cttresac	6.11	2.26	6	10 (0-10)	-0.47
grdfincc	4.14	2.33	4	10 (0-10)	0.09
viapolc	3.47	2.11	3	10 (0-10)	0.41
keydecc	4.21	2.15	4	10 (0-10)	0.21
implvdm	9.16	1.30	10	10 (0-10)	-2.24
panclobo	6.97	2.60	7	10 (0-10)	-0.65
panresmo	6.04	2.73	6	10 (0-10)	-0.41
gvhanc19	4.56	2.25	5	10 (0-10)	-0.02
gvjobc19	3.92	2.26	4	10 (0-10)	0.22
gveldc19	4.25	2.26	4	10 (0-10)	0.10
gvfamc19	4.52	2.14	5	10 (0-10)	0.04
hscopc19	5.14	2.26	5	10 (0-10)	-0.22
gvbalc19	5.02	2.37	5	10 (0-10)	0.03
gvimpc19	4.44	2.27	4	10 (0-10)	0.08
scltrst	13.46	4.79	14	29 (1-30)	0.23

Table 4.4: Descriptive statistics of the numeric variables, we treat scale variables ranging from 0 to 10 as numeric. We present mean, standard deviation, median, range and skewness.

Model 1			
Predictors	y		
	Odds Ratios	CI	p
(Intercept)	0.24	0.02 – 2.24	0.215
trstlgl	0.92	0.86 – 0.98	0.014
trstplt	0.87	0.81 – 0.94	<0.001
stfedu	1.07	0.98 – 1.17	0.130
stfhlth	1.08	1.00 – 1.16	0.047
medcrgv	1.11	1.01 – 1.24	0.042
votedir	1.19	1.10 – 1.30	<0.001
gptpelc	0.80	0.72 – 0.88	<0.001
grdfinc	0.92	0.84 – 1.00	0.049
fairelcc	0.91	0.85 – 0.98	0.007
dfprtalc	0.94	0.89 – 1.00	0.063
rghmgrpr	1.07	1.01 – 1.14	0.018
votedirc	0.95	0.90 – 1.01	0.111
gptpelcc	0.90	0.85 – 0.96	0.001
gvctzpv	1.07	1.00 – 1.15	0.061
wpestop	0.92	0.86 – 0.99	0.029
accalaw	0.89	0.85 – 0.93	<0.001
gveldc19	1.14	1.04 – 1.25	0.004



gvfamc19	0.84	0.77 – 0.93	0.001
gvimpc19	1.11	1.02 – 1.21	0.011
netusoft	1.10	1.00 – 1.22	0.053
polintr	1.23	1.05 – 1.44	0.009
psppipla	1.12	0.97 – 1.29	0.120
lrscale	0.90	0.84 – 0.97	0.004
freehms	1.22	1.08 – 1.39	0.002
hmsacld	0.88	0.79 – 0.99	0.034
loylead	0.87	0.77 – 1.00	0.044
imsmetr	1.20	1.05 – 1.37	0.009
hinctnta	0.93	0.87 – 1.01	0.071
ipeqopt	0.87	0.76 – 1.00	0.057
impsafe	0.90	0.79 – 1.02	0.094
ipfrule	1.11	1.01 – 1.21	0.029
ipmodst	0.93	0.84 – 1.02	0.132
impfree	1.10	0.98 – 1.23	0.090
imprad	0.86	0.76 – 0.97	0.016
gvconcl9	1.15	1.02 – 1.30	0.020
poleng	0.79	0.69 – 0.89	<0.001
crmvct [2]	1.37	0.94 – 2.04	0.110
dscrgrp [2]	4.56	2.00 – 11.93	0.001
ctzcntr [2]	2.89	1.14 – 6.95	0.020
cnthum [2]	2.51	0.69 – 10.85	0.182
cnthum [3]	2.07	0.64 – 8.28	0.256
cnthum [4]	2.02	0.62 – 8.11	0.273
cnthum [5]	3.04	0.92 – 12.32	0.088
cnthum [55]	6.66	1.42 – 35.39	0.020
domicil [2]	0.47	0.31 – 0.69	<0.001
domicil [3]	1.00	0.72 – 1.40	0.980
domicil [4]	1.33	0.99 – 1.78	0.056
hhmmb1 [2]	0.74	0.57 – 0.96	0.022
edrec [2]	1.76	1.12 – 2.73	0.013
edrec [3]	0.78	0.55 – 1.10	0.156
edrec [4]	1.48	1.12 – 1.95	0.005
Observations	2327		
R ² Tjur	0.197		

Table 6.5: Model 1 - assume ordinal as continuous – LASSO- stepwise

Model 2			
<i>Predictors</i>	<i>y</i>		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.19	0.02 – 1.76	0.147
trstlgl	0.91	0.85 – 0.97	0.004
trstplt	0.90	0.84 – 0.96	0.002
stflife	0.94	0.88 – 1.01	0.112
stflhth	1.09	1.02 – 1.16	0.010
atchctr	1.07	0.98 – 1.18	0.149
atcherp	0.96	0.90 – 1.02	0.147
medergv	1.09	0.99 – 1.21	0.087
votedir	1.19	1.10 – 1.29	<0.001
gtpelc	0.79	0.71 – 0.87	<0.001
fairelcc	0.90	0.84 – 0.96	0.003
dfprtalc	0.94	0.88 – 1.00	0.035
rghmgprc	1.09	1.03 – 1.16	0.003
votedirc	0.95	0.89 – 1.01	0.101
gtpelcc	0.90	0.85 – 0.96	<0.001
gvctzpv	1.10	1.02 – 1.18	0.011



wpestopc	0.93	0.86 – 0.99	0.031
accalaw	0.89	0.85 – 0.93	<0.001
agea	0.99	0.98 – 1.00	0.087
gvimpc19	1.07	1.00 – 1.15	0.045
polintr	1.18	1.01 – 1.36	0.033
lrscale	0.91	0.85 – 0.98	0.014
gincdif	1.16	1.00 – 1.34	0.046
freehms	1.27	1.13 – 1.44	<0.001
hmsaold	0.87	0.77 – 0.97	0.013
loylead	0.88	0.77 – 1.00	0.052
imsmetn	1.20	1.05 – 1.38	0.008
ipeqopt	0.87	0.76 – 0.99	0.043
impdiff	1.09	0.99 – 1.20	0.085
ipfrule	1.09	1.00 – 1.19	0.059
ipmodst	0.91	0.82 – 1.00	0.058
iphlppl	0.88	0.77 – 1.01	0.075
gvconcl9	1.17	1.03 – 1.32	0.012
poleng	0.80	0.71 – 0.91	0.001
crmvct [2]	1.36	0.93 – 2.04	0.123
dscrgrp [2]	4.27	1.88 – 11.21	0.001
ctzcntr [2]	3.87	1.50 – 9.41	0.004
cnthum [2]	2.95	0.81 – 12.85	0.120
cnthum [3]	2.32	0.71 – 9.33	0.194
cnthum [4]	2.41	0.73 – 9.73	0.175
cnthum [5]	3.60	1.07 – 14.74	0.051
cnthum [55]	7.37	1.56 – 39.30	0.014
vteurmb [2]	0.49	0.31 – 0.76	0.002
vteurmb [33]	1.33	0.67 – 2.56	0.400
vteurmb [44]	1.59	0.41 – 5.10	0.457
vteurmb [55]	1.14	0.66 – 1.94	0.626
domicil [2]	0.49	0.32 – 0.74	0.001
domicil [3]	0.98	0.71 – 1.36	0.920
domicil [4]	1.40	1.05 – 1.87	0.021
hhmbl [2]	0.75	0.58 – 0.95	0.018
edrec [2]	1.86	1.18 – 2.89	0.007
edrec [3]	0.70	0.50 – 0.98	0.038
edrec [4]	1.51	1.15 – 1.99	0.003
Observations	2327		
R ² Tjur	0.192		

Table 6.6: Model 2 - ordinal as continuous - group lasso- stepwise

Model 3			
Predictors	y		
	Odds Ratios	CI	p
(Intercept)	0.06	0.00 – 0.86	0.040
trstlgl	0.90	0.84 – 0.97	0.003
trstplt	0.90	0.83 – 0.96	0.004
stfedu	1.08	0.99 – 1.18	0.101
stfhlth	1.08	1.00 – 1.16	0.059
atchctr	1.11	1.01 – 1.22	0.038
votedir	1.20	1.11 – 1.31	<0.001
gtpelc	0.81	0.74 – 0.89	<0.001
fairelcc	0.89	0.83 – 0.95	0.001
dfprtalc	0.91	0.86 – 0.98	0.008
rgmgprc	1.08	1.01 – 1.14	0.019
gtpelcc	0.89	0.84 – 0.95	<0.001
gvctzpcv	1.09	1.01 – 1.18	0.024

wpestopc	0.90	0.83 – 0.96	0.003
accalaw	0.89	0.84 – 0.93	<0.001
hinctnta	0.92	0.85 – 1.00	0.047
panfolru	0.94	0.88 – 1.00	0.038
gveldc19	1.11	1.01 – 1.22	0.029
gvfamc19	0.85	0.76 – 0.93	0.001
gvimpc19	1.10	1.01 – 1.20	0.034
polintr [2]	1.41	0.76 – 2.75	0.297
polintr [3]	2.62	1.42 – 5.09	0.003
polintr [4]	2.17	1.16 – 4.30	0.020
lrscale [1]	0.38	0.08 – 1.77	0.209
lrscale [10]	0.31	0.06 – 1.49	0.137
lrscale [2]	0.54	0.14 – 2.26	0.388
lrscale [3]	0.48	0.12 – 1.95	0.292
lrscale [4]	0.56	0.15 – 2.28	0.407
lrscale [5]	1.20	0.32 – 4.76	0.784
lrscale [6]	0.50	0.13 – 2.08	0.329
lrscale [7]	0.22	0.06 – 0.92	0.032
lrscale [8]	0.22	0.05 – 0.94	0.036
lrscale [9]	0.23	0.04 – 1.19	0.080
gincdif [2]	1.42	1.06 – 1.90	0.017
gincdif [3]	1.56	1.05 – 2.32	0.027
gincdif [4]	1.24	0.63 – 2.37	0.531
gincdif [5]	0.26	0.05 – 0.93	0.060
freehms [2]	1.70	1.21 – 2.40	0.002
freehms [3]	2.51	1.65 – 3.84	<0.001
freehms [4]	2.10	1.22 – 3.58	0.007
freehms [5]	2.69	1.48 – 4.85	0.001
hmsaold [2]	0.69	0.37 – 1.32	0.252
hmsaold [3]	0.94	0.51 – 1.78	0.849
hmsaold [4]	0.69	0.36 – 1.31	0.247
hmsaold [5]	0.58	0.31 – 1.09	0.088
loylead [2]	0.44	0.19 – 1.06	0.058
loylead [3]	0.48	0.22 – 1.14	0.085
loylead [4]	0.36	0.16 – 0.87	0.019
loylead [5]	0.29	0.13 – 0.70	0.004
cnthum [2]	2.39	0.62 – 11.19	0.230
cnthum [3]	2.24	0.64 – 9.86	0.242
cnthum [4]	2.10	0.60 – 9.27	0.281
cnthum [5]	3.17	0.89 – 14.14	0.098
cnthum [55]	7.52	1.39 – 46.76	0.023
ipfrule [2]	1.96	1.12 – 3.53	0.021
ipfrule [3]	1.42	0.81 – 2.58	0.234
ipfrule [4]	1.34	0.74 – 2.47	0.344
ipfrule [5]	2.11	1.17 – 3.91	0.015
ipfrule [6]	2.44	1.16 – 5.16	0.019
ipudrst [2]	0.75	0.50 – 1.15	0.181
ipudrst [3]	0.53	0.34 – 0.81	0.004
ipudrst [4]	0.59	0.36 – 0.96	0.034
ipudrst [5]	0.49	0.28 – 0.85	0.012
ipudrst [6]	1.59	0.65 – 3.84	0.307
gvconcl9 [2]	2.55	1.27 – 5.39	0.011
gvconcl9 [3]	2.87	1.44 – 6.06	0.004
gvconcl9 [4]	2.94	1.46 – 6.24	0.003
gvconcl9 [5]	2.57	1.24 – 5.61	0.014
crmvct [2]	1.43	0.96 – 2.15	0.085
dscrgrp [2]	5.66	2.49 – 14.83	<0.001



ctzcntr [2]	2.76	1.07 – 6.80	0.030
domicil [2]	0.44	0.29 – 0.67	<0.001
domicil [3]	0.84	0.59 – 1.19	0.327
domicil [4]	1.22	0.91 – 1.64	0.182
mbtru [2]	2.07	1.02 – 4.31	0.046
mbtru [3]	1.71	0.96 – 3.19	0.079
hhmmb1 [2]	0.70	0.53 – 0.91	0.008
edrec [2]	1.33	0.83 – 2.11	0.223
edrec [3]	0.78	0.56 – 1.07	0.121
edrec [4]	1.42	1.06 – 1.89	0.017
Observations	2327		
R ² Tjur	0.247		

Table 6.7: Model 3 - ordinal as factors - lasso- stepwise

Model 4			
Predictors	y		
	Odds Ratios	CI	p
(Intercept)	0.33	0.02 – 4.31	0.402
trstprl	0.93	0.85 – 1.00	0.064
trstlgl	0.90	0.84 – 0.97	0.005
trstplt	0.92	0.84 – 1.00	0.042
stfgov	1.11	1.01 – 1.21	0.022
stfhlth	1.10	1.02 – 1.18	0.009
atcherp	0.96	0.90 – 1.02	0.226
votedir	1.23	1.13 – 1.33	<0.001
gtpelc	0.81	0.73 – 0.89	<0.001
fairelcc	0.90	0.84 – 0.97	0.004
dfprtalc	0.92	0.86 – 0.98	0.010
rghmgprc	1.11	1.04 – 1.18	0.001
votedirc	0.94	0.89 – 1.01	0.073
gtpelcc	0.89	0.84 – 0.95	<0.001
gvctzpc	1.12	1.04 – 1.21	0.004
wpestopc	0.91	0.84 – 0.98	0.019
accalaw	0.89	0.84 – 0.93	<0.001
hinctnta	0.91	0.84 – 0.98	0.014
panfolru	0.93	0.87 – 0.98	0.014
polintr [2]	1.35	0.70 – 2.72	0.390
polintr [3]	2.49	1.29 – 5.05	0.009
polintr [4]	1.76	0.87 – 3.71	0.125
cptppola [2]	0.81	0.57 – 1.14	0.220
cptppola [3]	1.14	0.78 – 1.65	0.503
cptppola [4]	0.65	0.38 – 1.10	0.115
cptppola [5]	0.50	0.16 – 1.43	0.217
lrscale [1]	0.38	0.08 – 1.82	0.221
lrscale [10]	0.19	0.04 – 0.95	0.041
lrscale [2]	0.43	0.10 – 1.79	0.232
lrscale [3]	0.43	0.11 – 1.75	0.226
lrscale [4]	0.44	0.11 – 1.80	0.241
lrscale [5]	0.92	0.24 – 3.66	0.905
lrscale [6]	0.41	0.10 – 1.71	0.210
lrscale [7]	0.16	0.04 – 0.68	0.011
lrscale [8]	0.16	0.04 – 0.68	0.011
lrscale [9]	0.17	0.03 – 0.90	0.038
gincdif [2]	1.40	1.05 – 1.88	0.023
gincdif [3]	1.57	1.05 – 2.34	0.028
gincdif [4]	1.31	0.66 – 2.51	0.432
gincdif [5]	0.27	0.05 – 1.03	0.082



freehms [2]	1.71	1.23 – 2.37	0.001
freehms [3]	2.52	1.68 – 3.78	<0.001
freehms [4]	2.05	1.21 – 3.44	0.007
freehms [5]	2.90	1.61 – 5.15	<0.001
loylead [2]	0.38	0.16 – 0.95	0.031
loylead [3]	0.45	0.20 – 1.09	0.064
loylead [4]	0.33	0.14 – 0.81	0.011
loylead [5]	0.27	0.12 – 0.67	0.003
cnthum [2]	3.34	0.80 – 16.78	0.117
cnthum [3]	3.51	0.92 – 16.62	0.086
cnthum [4]	3.66	0.95 – 17.45	0.077
cnthum [5]	5.62	1.44 – 27.08	0.020
cnthum [55]	13.88	2.33 – 94.24	0.005
ipshabt [2]	0.82	0.55 – 1.21	0.317
ipshabt [3]	0.83	0.55 – 1.27	0.395
ipshabt [4]	1.07	0.68 – 1.70	0.758
ipshabt [5]	0.57	0.34 – 0.95	0.032
ipshabt [6]	1.21	0.60 – 2.42	0.587
ipfrule [2]	2.25	1.26 – 4.13	0.008
ipfrule [3]	1.62	0.90 – 3.01	0.113
ipfrule [4]	1.62	0.88 – 3.06	0.132
ipfrule [5]	2.45	1.34 – 4.62	0.005
ipfrule [6]	3.28	1.52 – 7.15	0.003
ipudrst [2]	0.84	0.55 – 1.29	0.416
ipudrst [3]	0.60	0.39 – 0.95	0.027
ipudrst [4]	0.64	0.38 – 1.06	0.081
ipudrst [5]	0.50	0.27 – 0.90	0.021
ipudrst [6]	1.58	0.61 – 4.05	0.339
ipgdtim [2]	0.57	0.37 – 0.89	0.012
ipgdtim [3]	0.49	0.31 – 0.77	0.002
ipgdtim [4]	0.61	0.37 – 0.98	0.042
ipgdtim [5]	0.67	0.40 – 1.13	0.129
ipgdtim [6]	0.73	0.33 – 1.57	0.431
gvconc19 [2]	2.40	1.18 – 5.13	0.018
gvconc19 [3]	2.83	1.41 – 5.99	0.005
gvconc19 [4]	3.18	1.58 – 6.74	0.002
gvconc19 [5]	2.71	1.30 – 5.92	0.010
crmvct [2]	1.56	1.04 – 2.36	0.034
dscrgrp [2]	4.58	1.99 – 12.12	0.001
ctzcntr [2]	2.40	0.91 – 6.03	0.066
vteurmb [2]	0.57	0.35 – 0.91	0.020
vteurmb [33]	1.56	0.75 – 3.14	0.217
vteurmb [44]	1.85	0.50 – 6.09	0.327
vteurmb [55]	1.37	0.76 – 2.45	0.288
gndr [2]	1.19	0.93 – 1.53	0.165
domicil [2]	0.46	0.30 – 0.70	<0.001
domicil [3]	0.89	0.62 – 1.26	0.501
domicil [4]	1.25	0.92 – 1.69	0.153
hhmmb1 [2]	0.69	0.53 – 0.90	0.007
edrec [2]	1.43	0.89 – 2.26	0.132
edrec [3]	0.76	0.55 – 1.06	0.109
edrec [4]	1.50	1.12 – 2.01	0.007
Observations	2327		
R ² Tjur	0.258		

Table 6.8: Model 4 - ordered as factors- group lasso – stepwise

List of figures

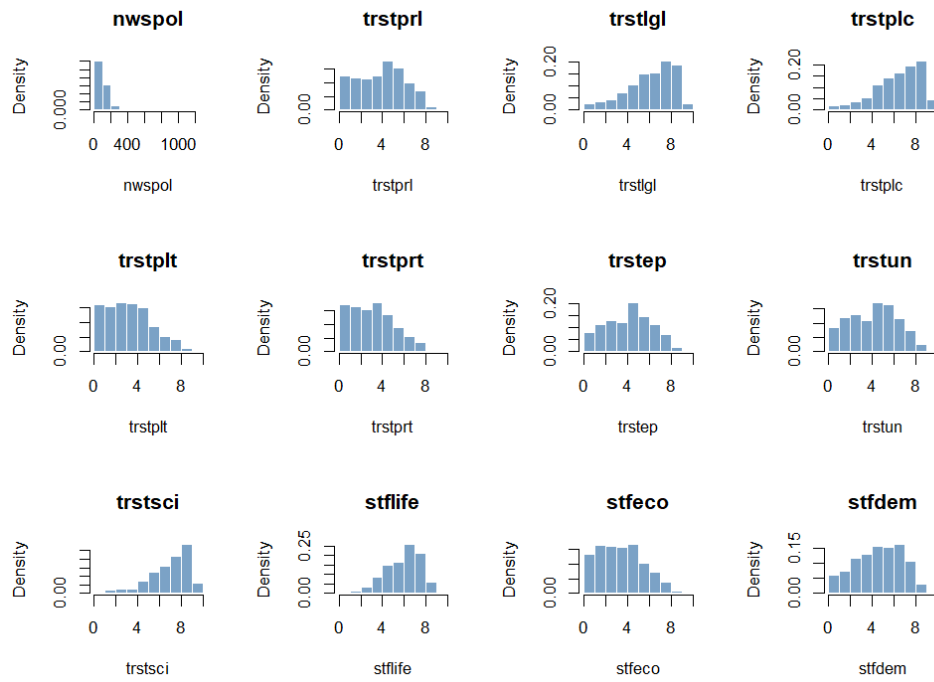


Figure 4.8: Histograms of independent variables treated as numeric 1/4

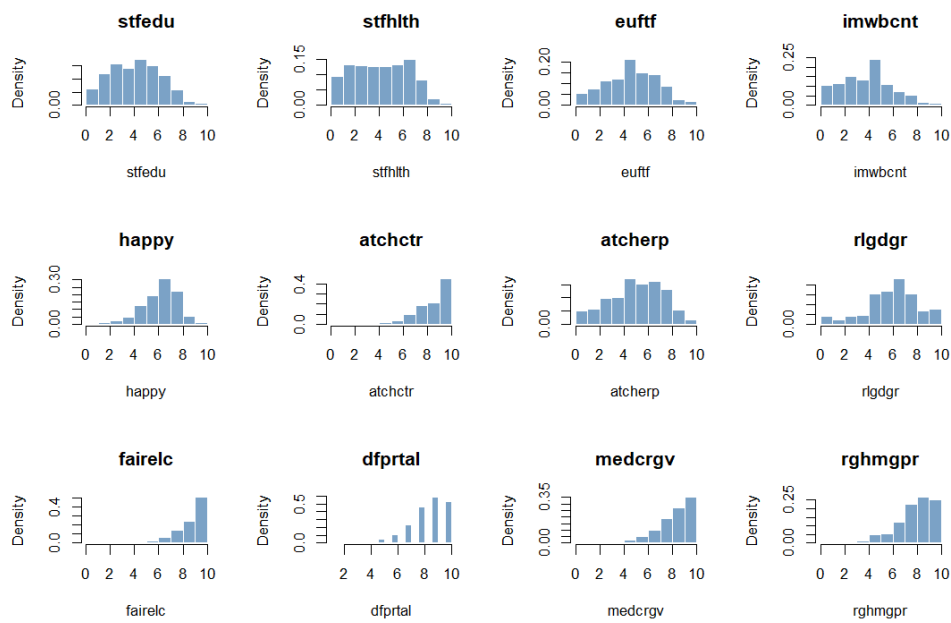


Figure 4.9: Histograms of independent variables treated as numeric 2/4



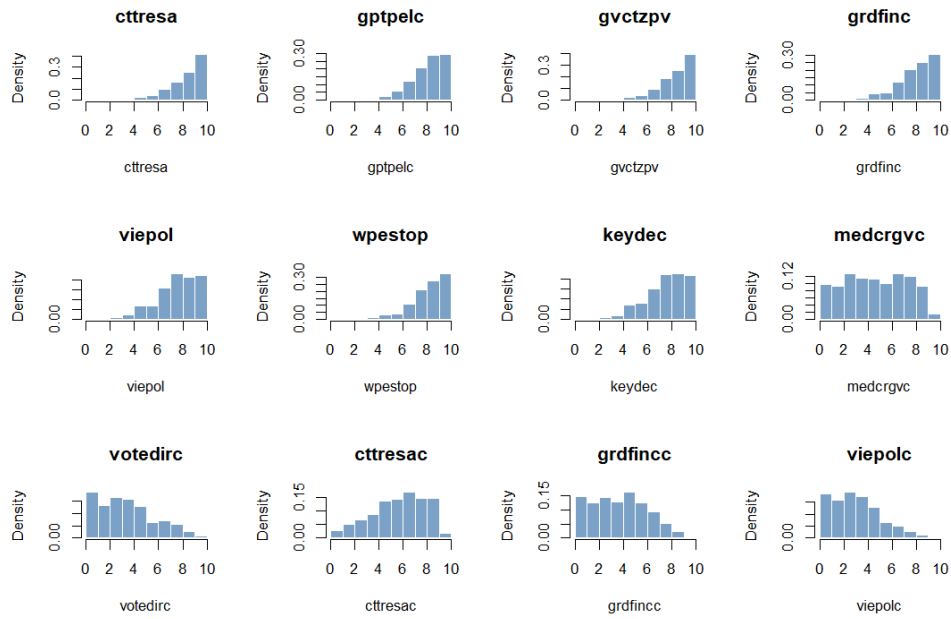


Figure 4.10: Histograms of independent variables treated as numeric 3/4

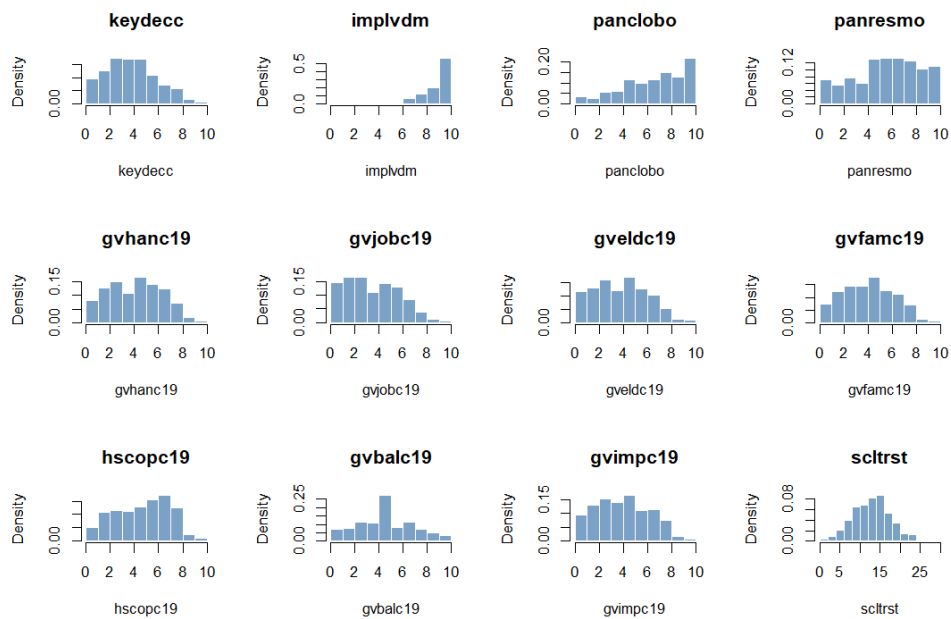


Figure 4.11: Histograms of independent variables treated as numeric. 4/4



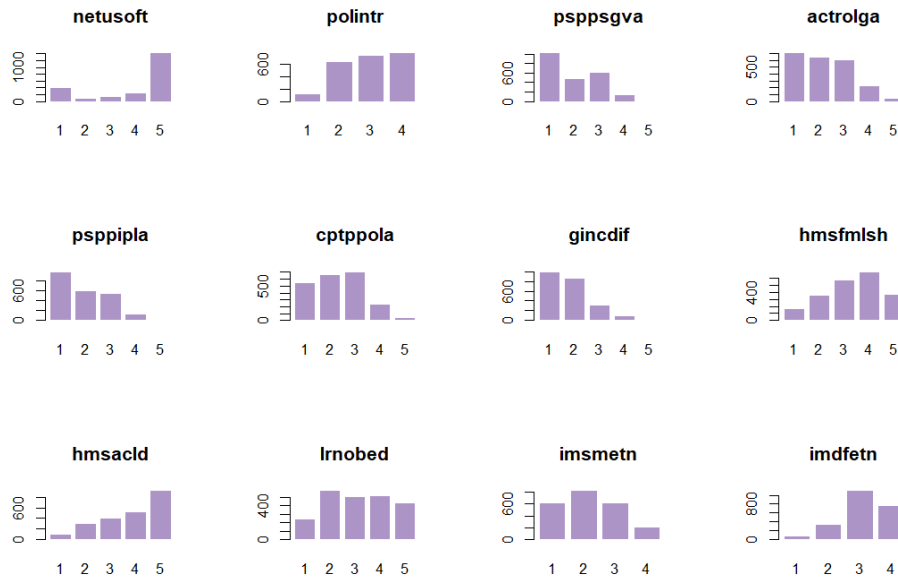


Figure 4.12: Bar plots for independent variables treated as ordered factors.1/4



Figure 4.13: Bar plots for independent variables treated as ordered factors.2/4



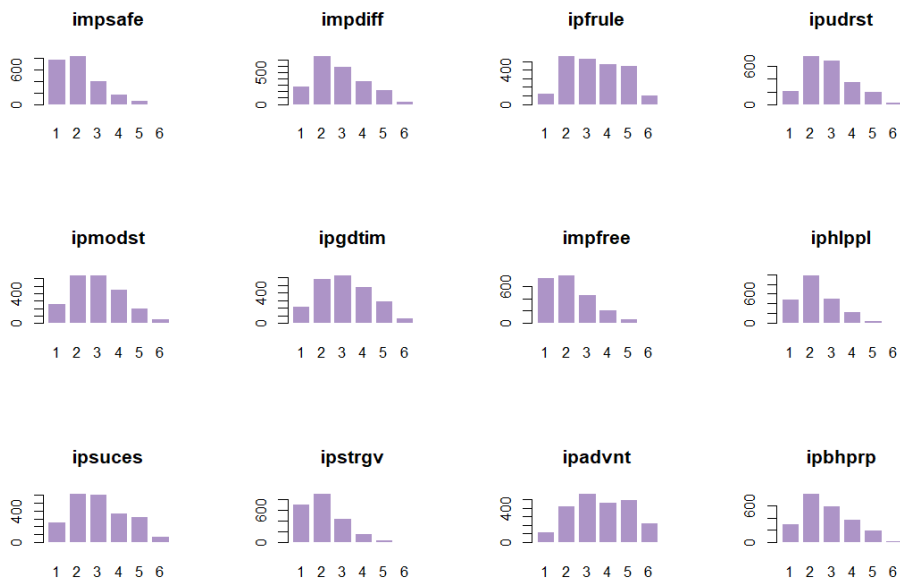


Figure 4.14: Bar plots for independent variables treated as ordered factors.3/4

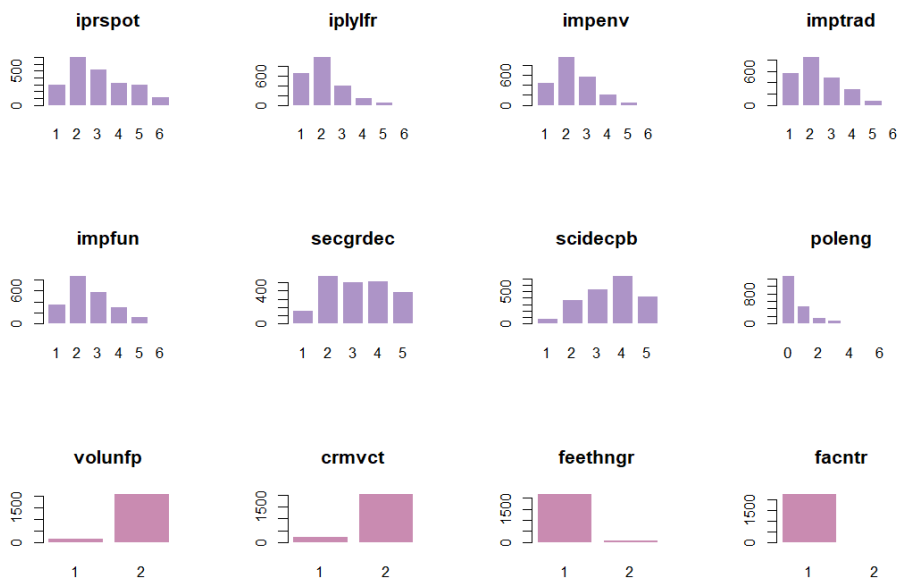


Figure 4.15: Bar plots for independent variables treated as ordered factors and as factors.4/4



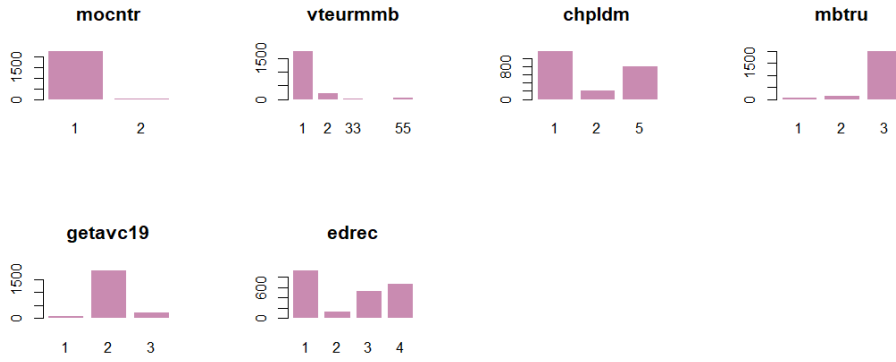


Figure 4.16: Bar plots for independent variables treated as ordered factors.

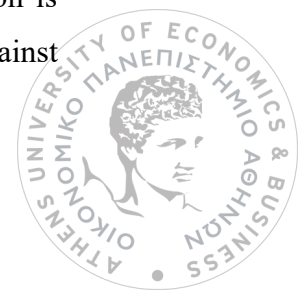


An Overview of the Political Parties in Greece

In order to fully understand the dynamics of the survey data and interpret our findings correctly, it is essential to consider the political environment in Greece during the time of data collection. The data were collected from November 2021 to May 2022. The variable we are analyzing is referring to the question “What did you vote for in the last parliamentary election (2019)?”. We provide an overview of the main political parties that shaped public opinion and potentially influenced survey responses.

An adequate understanding of the Greek political parties that participated in the July 2019 parliamentary elections is essential for our analysis. We will outline below some characteristics to form a picture of the ideology and political stance of the parties and, by extension, their voters. The information we present here comes from the publication of Elias et al (2022) *The Greek political landscape 2019-2021*. This publication emphasizes the key features of Greece's political and party system. Its primary goal is to analyze the structure of the Greek political landscape following the 2019 parliamentary elections and in the context of the pandemic crisis. They use graphs to illustrate each party's ideological stance within a two-dimensional political space, based on their stances on 35 key policy issues in current public debate.

- New Democracy (Νέα Δημοκρατία – ND): It is the biggest center right party in the Third Greek Republic founded by Konstantinos Karamanlis after the fall of the dictatorship in 1974. New Democracy portrays itself as a modern center-right political party that emphasizes economic reforms and managerial competence. The party seeks to limit state intervention in the economy and is open to further flexibilization of labor rights. During 2019 – 2023 the government of New Democracy promoted foreign investment. Moreover, it prioritizes public university reform, including the abolition of academic asylum and allowing private universities. ND is strongly pro-EU and takes a tough stance on law and order, as seen with the law that banned the manifestations in 2020. In healthcare, ND supports a predominantly public system supplemented by private initiatives. Its approach to religious affairs during the pandemic was mixed, and it has shown some openness to symbolic inclusivity by appointing women and an openly gay politician to government roles, balancing conservative and progressive elements. The party's stance on immigration is restrictive, they expand camps on specific islands, and they passed laws against



refugee's rights, as already mentioned. ND tends to avoid significant changes on moral and identity issues, focusing on a strategy to appeal across the political spectrum.

- SYRIZA (Coalition of Radical Left – ΣΥΡΙΖΑ): SYRIZA has evolved from a marginal left-wing alliance to a significant political force in Greece. It gained power by promising to free the country from austerity measures. SYRIZA remains a key player, particularly among younger voters. Economically, SYRIZA focuses on wealth redistribution and environmental sustainability and protecting workers' and middle-class rights. The party advocates for progressive policies on immigration, multiculturalism, gender equality, and LGBTQI+ rights.
- KINAL (former PASOK, Movement of Change – Panhellenic Socialist Movement – ΠΑΣΟΚ): KINAL, previously known as PASOK, has faced significant challenges since the financial crisis of 2010. It used to be a dominant force in Greek politics with a legacy of major social and political reforms, but the party struggled with scandals and financial crises which led to its decline. The party advocates for state intervention in the economy, worker protection, and a strong welfare state. It supports progressive socio-cultural issues such as gender equality, LGBTQI+ rights, and an open society, while maintaining a nuanced stance on immigration. KINAL is pro-European and it favors EU integration and policies, but prioritizes national interests in defense and foreign policy.
- KKE (Communist Party of Greece): The Communist Party of Greece (KKE), founded in 1918, is the oldest political party in Greece. It is known for its communist principles. KKE advocates for a socialist system, emphasizing class struggle and rejecting coalition politics with other progressive parties. The party maintains a hard Eurosceptic stance, calling for Greece's exit from the EU and NATO, and focuses on representing the working class through policies promoting labor protection, social security, and public education. Despite electoral challenges, KKE has remained a stable force in Greek politics by appealing to its historical and ideological traditions and forming strong relations with unions and syndicates.



- **Elliniki Lisi (Greek Solution):** Greek Solution (EL), founded in 2016 by Kyriakos Velopoulos, emerged as a prominent populist radical right party in Greece after replacing the neo-Nazi Golden Dawn in 2019. The party advocates for conservative cultural values, and a centrally planned economy. It opposes EU interference, supports increased defense spending, and holds strict views on immigration and LGBTQI+ rights. Velopoulos, leveraging his TV persona, remains the party's central figure, driving its agenda and maintaining its visibility.
- **MERA25 (Diem25):** MeRA25, founded in 2018 by former Finance Minister Yanis Varoufakis, is a leftist party with a strong anti-austerity stance, opposing bailout agreements and advocating for public debt restructuring. The party emphasizes economic reform, including a "European Green New Deal," and demands radical changes in EU institutions for greater transparency and accountability. Varoufakis's leadership and popularity are central to MeRA25, which promotes multiculturalism and progressive social policies while maintaining a critical view of EU integration.



