JOBS FOR GERMANS?: THE IMPACT OF UNEMPLOYMENT ON AFD SUPPORT

BY EYLÜL BEGÜM SAĞLAM

December 18, 2024

Abstract: As the recent election opinion polls in Germany reflect, the prominent right-wing populist Alternative für Deutschland (AfD) takes the lead as the second party following Christlich Demokratische Union Deutschlands (CDU), by leaving all other mainstream parties behind (Clarke 2024). Moving from Germany and AfD, this pre-analysis plan focuses on the debate in the populism literature on what motivates support for right-wing populist parties. Using the most recent data (round 11) published by European Social Survey (ESS), I analyze whether unemployment increases the political support for right-wing populist parties by focusing on the case of Germany and Alternative für Deutschland (AfD). Through employing matching and logistic regression, I argue that unemployment, a component of economic grievances camp in the literature, provide explanatory reasons for why AfD has become a prominent party in German politics. The findings of this research are expected to contribute to the literature on why economic factors are essential to understand support for right-wing populist parties in Germany.

1 Substantive Question

As[1] the right-wing populist parties show more presence in parliamentary politics all around Europe, it becomes a vital question for social scientists to analyze the mechanism behind increasing support for these parties and to interpret demands of voters from the system. The "grievance theory" in the literature focuses on economic and socio-cultural grievances faced by the losers of globalization as drivers of support for right-wing populist parties (Golder 2016; Bernburg 2015; Grasso and Giugni 2016; Kern, Marien, and Hooghe 2015; Rüdig and Karyotis 2014; Kurer et al. 2019). On the one hand, the literature argues that the populist party support could be explained better by "economic grievances," such as unemployment (Golder 2016). On the other hand, however, scholars highlight that it is not necessarily the economy, but factors related to socio-cultural grievances, such as immigration, that constitute the most essential reason for the increasing vote share of populist parties (Schwander and Manow 2017).

This pre-analysis plan mainly aims to analyze the argument of the former camp on economic grievances, to see if issues like unemployment may be driving mechanism for individuals' support for right wing populist parties while adjusting for socio-economic grievances. Accordingly, I analyze the following research question as a part of this paper: Does experience with prolonged unemployment experience increase individuals' vote support for right-wing populist parties such as the AfD?

2 Hypothesis

I argue in this pre-analysis plan that having experienced unemployment for a while at some point in their lives influence people's support for right-wing populist parties in a positive way by moving from the case in Alternative für Deutschland (AfD) in Germany. The extent of unemployment in this context is defined as being unemployed and actively seeking work for more than three months in some time of their lives. Based on the "grievance theory," I aim to analyze if economic problems faced by individuals argued to be influential to increase their discontent with mainstream parties

[1] This pre-analysis plan answers sixteen questions based on the final template provided by Bowers (2024). The titles of each section is mainly organized based on the structure of EGAP Pre-Analysis Plan Chen and Grady (2024).

and precipitate them to resort to parties which promise for substantial change in the status quo (Golder 2016; Bernburg 2015; Grasso and Giugni 2016; Kern, Marien, and Hooghe 2015; Rüdig and Karyotis 2014; Kurer et al. 2019).

Focusing on this question is meaningful due to three main reasons. Primarily, it sheds light on the political decisions driven by economic motivations. Secondly, question attempts to seek an explanation for a global trend, where many right-wing populist parties receive increasing support from electorates (Silver 2022). Finally, contributions of possible answers emerging from this research may expand beyond academia and breed policy suggestions which could mitigate employment and prevent social conflicts between different groups (e.g., citizens and immigrants) in societies.

Hypothesis 1. *Individuals who have experienced prolonged unemployment and actively sought work for more than three months are more likely to support the AfD in elections compared to those who have not, adjusting for five main covariates on education, age, gender, religion and perceptions about immigrants' impact on economy.*

3 Data and Research Design

In order to measure the relationship between unemployment and AfD support, I will follow an observational study design that uses quasi-experimental techniques such as matching methods. For this purpose, European Social Survey (ESS)'s latest data (Round 11) is a helpful source. This data set analyzes twenty-four European countries: Austria, Belgium, Croatia, Cyprus, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Lithuania, The Netherlands, Norway, Poland, Portugal, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland and the United Kingdom. The unit of analysis in the data set are individual respondents who participated in the survey. Since Germany is the focus of the analysis in this paper, data consists of 2420 individuals who provided responses to various survey questions by ESS. The data set is publicly available, and its respective cross-national module focuses mostly on gender and socio-economic inequalities via surveys conducted by a mixed-method approach (in-person interviews and in-person/online questionnaires). It has various measures related to demographics of individuals, their political views, employment situations and attitudes. The selected independent variable is whether individual ever had an experience with unemployment for more than three months, while dependent variable is if individual voted for AfD in the latest election. Covariates which will be subject to matching process are five main socio-demographic variables: age, gender, education level, religion, and perceptions about immigrants' impact on economy (i.e., a ranked scale on whether immigration is good or bad for country's economy.) I chose five main covariates to match on following the variables analyzed by Rooduijn (2018) in a cross-national study comparing similarities in supporters of populist parties, which also uses ESS as data.

Based on the goal of analyzing the impact of unemployment on AfD support, the design of this research will start with bipartite matching. I chose this strategy since I compare two separate groups (treatment and control on being unemployed and employed, respectively) while keeping other background covariates similar. Treatment units will only be matched with control units, and vice versa. I employ matching process on five covariates (age, gender, education level, religion, and perceptions about immigrants' impact on economy) via exact and propensity score matching to balance treatment and control groups. Since the respective survey responds consist of categorical or ranked variables [2], I chose this strategy to provide a robust balance in treatment and control groups. I will do an omnibus test as the next step to see if matching is balanced, following Hansen and Bowers (2008) I analyze standardized mean differences to see if groups are in a comparable shape after matching. To see the relationship between unemployment and AfD support, I use logistic regression which will be estimated on the matched data obtained via exact matching and propensity score matching. I chose logistic regression by following a similar study measuring predictors impacting right-wing populist party support in Europe (Rooduijn 2018). Logistic regression will be helpful in this context since both independent and dependent variables are coded as binary. After hypothesis testing and judgement on statistical estimation process, I conduct sensitivity analysis for unobserved variables to assess the robustness of the findings.

[2] As I will explain in the later sections of the paper, there is only one continuous variable (age), which is also arranged to be an ordered variable to make matching process less problematical (i.e., finding match with specific age could have been difficult).

4 Advantages and Disadvantages of Research Design

The research design is advantageous in this particular case, since the aim is to see unemployment's impact on AfD support, and matching enables isolating this treatment's effect based on similarities in observed variables such as religion, gender, age, education-level, and perceptions about immigrants' impact on economy. Along these lines, I aim to explain the differences in AfD support between unemployed and employed groups by unemployment (treatment effect). By balancing on covariates, matching also reduces my reliance on strong parametric assumptions such as normality before I apply regression for hypothesis testing purposes. This way, the hypothesis test is done on a balanced data, and the results of logistic regression analysis are less sensitive to bias and any model specification problems (e.g., adjusting, linearity assumption, interaction variables) that would not be easily resolved without a pre-matching balance. Seeing SMD values of covariates after omnibus test facilitates checking the distribution of treatment and control after matching process. Sensitivity analysis that follows after hypothesis testing and assessment of the performance of statistical estimations adds an additional step to test findings by focusing on unobserved covariates' impact. This will help if the relationship between unemployment and AfD support might be affected by some unobserved confounder(s).

The main disadvantage of the research design is that it is an observational study, therefore, it does not promise a strong causal inference process as would have been provided by randomization and randomized controlled trials (RCTs) even if it employs quasi-experimental techniques. The reason is that quasi-experimental techniques do not have "an explicit random assignment procedure," which makes their "causal inferences subject to greater uncertainty" (Rosenbaum, Rosenbaum, and Briskman (2010, 17). In addition, matching works with only five variables in this research. Therefore, even with a sensitivity analysis done afterwards, matching process may still face omitted variable bias, since it cannot fully explain the impact of unobserved variables. Thirdly, it is challenging to match in a balanced manner, since there are so many variables with different response structures even after they are re-classified in different groups (e.g., binary, ranked, categorical). Fourthly, part of the data may get lost after matching as some unmatched units are excluded in the matching process. Fifthly, some covariates are organized in more simplified categories (e.g., grouping education in three categories rather than more specified twenty-seven categories as it is done by ESS excluding missing data) to make matching process easier, but this is still not adequate to balance some covariates fully (e.g., religion) when number of units with particular groups (e.g., respondents identifying themselves as Christian) are more than others. In addition, further categorization of covariates for the sake of better balance as well as binary categorization of both dependent and independent variables may precipitate neglecting variations in categories (e.g., differences between Christian groups, generational discrepancies between a twenty-year old and a twenty-nine year old in the young adult category of age). It is also important to note that though this research aims to see the relationship between unemployment and AfD support, this is very specific to Germany case, and may not be applicable to understand right-wing populist party instances in other contexts.

5 Measures and Indices

I finalized the list of measures after a data cleaning process in the ESS website before downloading the data, where I chose respective variables that I plan to use in this research. I measure the primary outcome, the party voted in the latest elections, by using the variable coded in the data set as prtvgde2 which specifically asks respondents which party they voted for in the last national elections as a second vote in Germany context. The measure consists of ten categories in addition to missing values, but is recoded as a binary variable for the purposes of this paper: if individual respondent voted for AfD, the prtvgde2 measure takes the value 1. If respective respondent voted for another party, this measure takes value 0. The variable on second vote is deliberately chosen to capture support for party rather than individual candidate, since first votes are given for an individual constituency candidate, while second votes are for a party-list in the respective state (Federal Ministry of the Interior and Community, n.d.). The explanatory variable (coded as unemp3m in the data set) is assignment to treatment, and asks if respondents were ever unemployed and seeking work for a period of more than three months. Individuals who chose the option "yes" (being unemployed) are assigned to treatment and coded as 1, while individuals who selected the option "no" are assigned to control and coded as 0. For the matching process, five main background variables are selected: religion, age, gender, education, and perceptions about immigrants' impact on economy (i.e., immigration is bad/good for the economy, ranked from 0 to 10 in order). Due to varieties in categories in the original data set and for the ease of matching process and analysis, I re-coded variables in particular categories as either ordinal or categorical variables. For religion, various categories in Christianity are united in "Christian," and all other religions and belief systems in "Non-Christian." For age, 0-29 [3] are coded as young adults, 30-49 coded as adults, 50-64 coded as middle aged and above 65 is coded as seniors to capture minor changes between small age differences. For education, I followed the low, medium, and high categorization made on International Standard Classification of Education (ISCED) by Eurostat (European Commission, n.d.). For perceptions about immigrants' impact on economy, I used a ranked scale were 0-5 are coded as low-medium and 6-10 are coded as medium-high. I provided more clear explanation on how I coded variables in Table 1.

Variable	Type	Coding
Unemployment (uemp3m)(main DV)	Binary	1= Unemployed, 0 = Employed
AfD Support (prtvgde2)(IV)	Binary	1= Voted AfD, 0 = Voted Other
Religion (rlgdnade_ group))	Binary	1=Christian and 0=Non-Christian
Age (age_ group)	Ranked	Young, Adult, Middle Age, Senior
Education (edulvlb_ group)	Ranked	Low, Medium, High
Immigrants' Impact on Economy (imbgeco_ group)	Ranked	Low-Medium and Medium-High
Gender (gndr)	Binary	1=Male, 0=Female

Table 1: Measures

6 Strategies

6.1 Identification and Adjustment Strategy

The main identification strategy employed in this research is the use of observational data to make inferences on the possible causal link between unemployment and AfD support. I use matching as the adjustment strategy and quasi-experimental design by employing exact matching on particular variables and propensity score matching with 0.2 caliper for the remaining ones. The treatment

[3] There are respondents under the age of 18, including respondents aged 16 in the ESS dataset.

is not random, but it is adjusted via matching and following logistic regression to make it "as-if randomized" as much as possible. With the help of optmatch package, I used exact matching on gender and religion variables. Exact matching strategy on these variables is useful, since they signal considerable baseline differences (e.g., unlike miniscule educational differences between a middle school graduate and second-year middle school student) between categories. Due to their ranked structure, I employed propensity score matching with 0.2 caliper for three covariate variables, age, education-level and economic attitudes toward immigrants. This mixed approach facilitated a stricter criteria on primary variables of interest while accommodating for other ranked variables. My goal is to create a balanced matching without having huge number of categories (e.g., trying to match on only three educational categories instead of twenty-seven categories in the ESS data) creating a problem for the matching process while not neglecting the impact of clear differences like being a male or identifying with particular religion.

After the matching process, I use logistic regression (via glm code in R) as an additional adjustment strategy for covariance imbalances. I use logistic regression as a second check mechanism to balance for residual discrepancies after matching is complete. I also used this regression process for hypothesis testing. Choosing logistic regression for testing provides flexibility to deal with many types of variables (binary, categorical, ordinal) and assumptions on normality with regard to how outcome variable is distributed. Since this research only balances for five covariates, there may be many factors which are not properly captured by the matched variables. Therefore, I employ sensitivity analysis to consider possible impact of unobserved confounding as the next step after assessing the performance of hypothesis testing process and estimators.

6.2 Evaluating the Success of Adjustment Strategy

The adjustment done via matching provides a justification for how research design mimics "as-if randomized" experiments by making control and treatment groups similar, so that confounding effect is minimized. After the matching process, I conduct omnibus test to see covariate balance (Hansen and Bowers 2008), and I calculate standardized mean differences (SMDs) for all covariates and their categories, to see if treatment and control groups are balanced after matching. Since SMD shows mean differences between matched treated and control groups, SMD values closer to zero are expected to show a better match, and all covariates satisfy this requirement. I use 0.01 as the threshold. Gender (SMD = 0.0000 for both categories) and religion (SMD = 0.000) variables demonstrate perfect balance. Education (SMD = 0.0000 for medium group, SMD = 0.01 for high group, SMD = -0.03 for low group) and perceptions about immigrants' impact on economy (SMD = 0.02 for low-medium and -0.02 for medium-high group) show a well-balance despite higher values of some groups based on reference value of 0.01. The mean differences for age show perfect balance (SMD = 0.000) except senior group (SMD = 0.01). A detailed visualization of SMD values of all covariates with their groups are demonstrated below in Figure1. The dashed line denotes perfect balance of respective variables between matched treatment and control groups. The effective sample size after matching is 719.2 with 2379 observations, which refers to the notable part of the original data set with 2420 observations.Since I lost large group of observations after matching, I want to make sure that my sample size is large enough to detect the expected effect following hypothesis testing. Therefore, to judge the performance of tests in terms of power, I will conduct a simulation-based power analysis with 1000 simulations for all three models. I keep alpha = 0.05 as the significance level, 0.2 as control group probability (baseline probability of AfD support for control group), and 0.1 as treatment effect (unemployment's effect on AfD support). The result of the simulation shows estimated power as 0.981 for the main model, 0.977 for bivariate model, and 0.988 for the model with missingness. All three models in my analysis perform very high power, indicating that they are robust to detect the treatment effect. The model with missingness has a slightly higher power compared to two other models. This is understandable, since adding missingness indicators to the model helps correct for biases that occur as a result of missing data.

7 Plans for Missing and Extreme Data

To recall how I mentioned raw data provided by ESS has many categories (i.e., twenty-seven categories for education), I want to indicate once again that I categorize variables in different

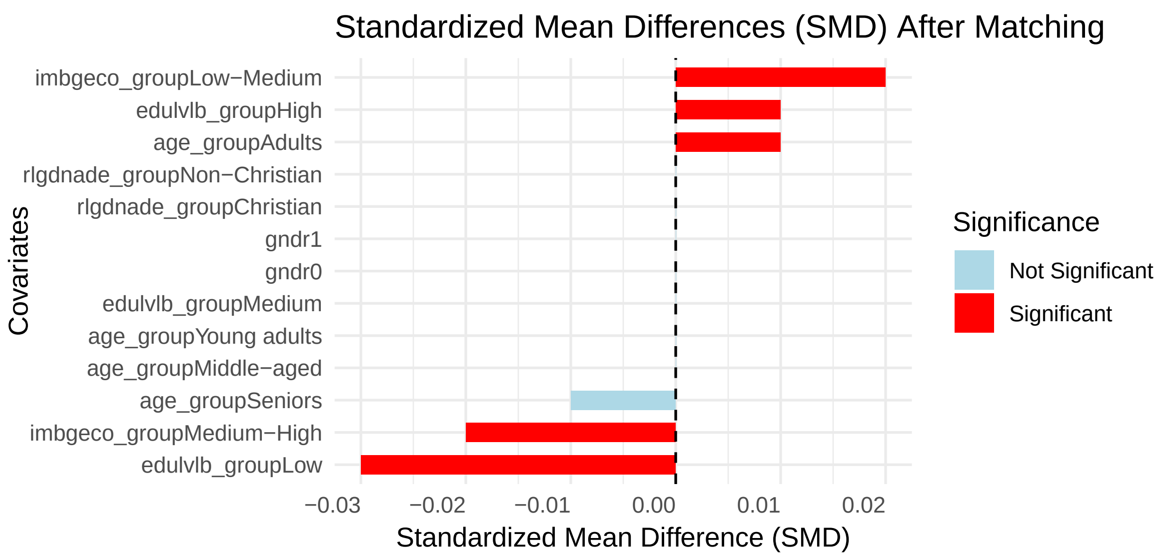

Figure 1: Omnibus Test on SMD Values

groups suitable for the matching process. The rationale for this strategy came from facilitating matching process while tolerate for miniscule differences between groups. It would unnecessarily complicate the matching process with many levels that do not demonstrate huge differences (e.g., the differences between completing first two years of high school and first three years of high school) between observations while making matching itself extremely difficult. These newly variables are added as separate columns to the original data set (e.g., age_ group, edulvlb_ group, imbgeco_ group, rlgdnade_ group). As the first step after this process, I used colSums(is.na(...)) code to check for missing values across all variables subject to my research. The variables which had missing values were outcome variable prtvgde2 (787), treatment variable uemp3m (6), and two of the covariates rlgdnade_ group (1169), imbgeco_ group (34) and edulvlb_ group (14). Two other covariates age_ group and gndr did not have any missing values.

After identifying the missing data, I decided to use two main strategies to deal with it. Firstly, I added missingness indicators (e.g., imbgeco_ group_ missingness) to check whether missingness was due to randomness or a deliberate pattern for all seven variables, inspired from Rabb et al. (2022). They are coded as 1 if there is missingness in the respective variable, 0 if there is no missingness. Missingness indicators will also be helpful for further regression processes to see if these missing variables constitute a confounder effect. By using VIM as library and ggplots for visualization

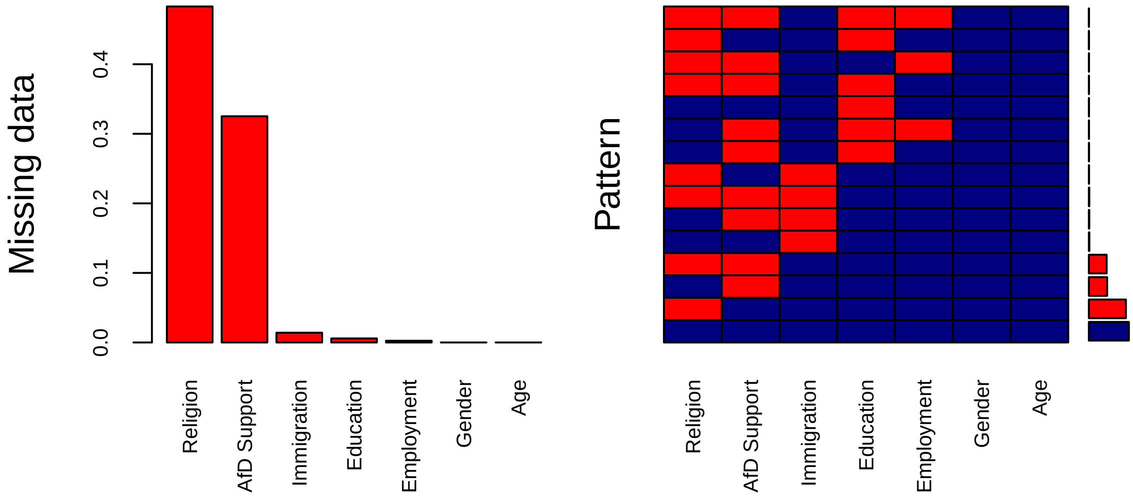

Figure 2: Missingness Pattern in Data

in Figure 2, I can interpret that especially in outcome variable on AfD Support (prtvgde2) and religion (rlgdnade_ group), there is considerable missingness. When missingness among variables are compared, it also does not reflect any pattern, since I cannot comment clearly on when and whether particular variables are always missing together. This shows that missingness may be due to randomness.

Based on this, I employ multiple imputation (via "mice" library) on covariates while dropping age and gender from the process, since these variables have zero missing values. I chose to include main explanatory and outcome variable in the imputation process alongside three other background variables edulvlb_ group, imbgeco_ group, and rlgdnade_ group. Including treatment and outcome variables in the imputation is risky due to the possible bias that may result by the impact of imputation process predictions on the relationship between main explanatory and outcome variable. However, since outcome variable has the highest number of missing values among all variables, I do not want to face a rapid decrease in the sample size. I did not specifically employ any actions for extreme values, since my covariates are either categorical or ordinal. Grouping them in categories in the earlier stages managed any possible problems related to extremeness for variables like age.

8 Statistical Tests and Decision Making Criteria

I choose to employ logistic regression based on Rooduijn (2018) who conducted a similar study by using binary outcome variable. I conduct this test on three models: the main model with five covariates and treatment variable; the bivariate model focusing only on treatment and outcome variable; and a model which includes variables in the main model and additionally, missingness indicators of these covariates. My rationale to add three models is to highlight how matching process before statistical tests matter to adjust for covariates, and whether adding missingness indicators considerably changes results. I evaluate these models based on their confidence intervals, odd ratios and AIC values for significance, effect size and model fit, respectively.

9 Performance of Statistical Tests

When I compare three models with one another, I can interpret that based on AIC values, the model with missingness indicators is the best fit since it has the lowest AIC value (918.65) compared to both bivariate (1036.5) and main model with five covariates (924.35). This shows that bivariate model does not account for any missingness or the impact of other covariates, while adding missingness improves model only slightly compared to the main model, as low coefficient values of missingness variables highlight. This may show that although adding a model with missingness indicators may be a good strategy for seeing the effect of missing observations, missingness does not necessarily change the model fit considerably in this context.

The main model includes unemployment (uemp3m), age group, gender (gndr), education level (edulvlb_ group), economic perceptions on immigration (imbgeco_ group), and religion (rlgdnade_ group). The confidence interval of odd ratios is used for analysis, and criteria for confidence interval is if the interval excludes 1, it shows a meaningful relationship, and values below 1 show a negative relationships while values above 1 show a positive relationship. Accordingly, I use confidence intervals to be able to see if there is an effect and how large it is (magnitude). The confidence interval for intercept (0.02 to 0.04) show the baseline levels without covariates, which is quite narrow to express the support for AfD. I can see that many covariates have considerable effects on AfD support. The unemployment variable shows a strong and positive relationship with AfD support, with the confidence interval between 1.46 and 3.07, indicating that unemployed individuals are expected to support AfD between 1.46 and 3.07 times more likely. Age covariate shows that for all categories, results show that the effect is ambiguous, since all three linear, (CI: 0.51-1.17), quadratic (i.e., effect accelarates as age decreases) (CI: 0.56-1.22) and cubic (i.e., effect becoming strong at lower ages) (CI: 0.88-1.72) components show that as age decreases, AfD support is also expected to increase, but none of the values provide meaningful relationship. Gender (males) also has a robust link with AfD suppport, with a confidence interval of 1.29 to 2.69, suggesting that males are expected to support AfD more likely than females. Education variable shows that those with higher education are expected to be more likely to support AfD, with a confidence interval of 1.23 to 2.34, while this effect increases at a decreasing rate according to quadratic case (CI: 0.35 to 0.94). Economic perceptions on immigrants also play a role, since individuals in the low-medium group are expected to be more likely to support AfD (CI: 0.20 to 0.39). Religion (being Christian) is expected to increase support for AfD, but displays no meaningful relationship, since confidence interval spans from 0.0035 to 0.29. This wide interval also shows less precision in the estimate.

The bivariate model in my analysis includes merely unemployment (uemp3m) as a predictor of AfD support. The confidence interval for unemployment (CI: 1.63 to 3.27) shows that being unemployed is expected to considerably increase the likelihood of supporting AfD, with the odds being 1.63 to 3.27 times higher for unemployed individuals. Nevertheless, since no other covariates are included, I need to note that the model does not account for the impact of other contextual factors like age, gender, or education.

Finally, in the model with missingness indicators, I can observe similar trends for the key covariates with main model. Unemployment still has a strong effect on AfD support, with the confidence interval being between 1.38 and 2.94, highlighting how unemployed are more 1.38 to 2.94 times more likely to support AfD. Age is not expected to have a meaningful impact on AfD support in this model, in linear (CI: 0.55 to 1.78), quadratic (C:0.55 to 1.20) and cubic (CI: 0.87 to 1.71). Gender (males) stays as a meaningful predictor with a confidence interval of 1.24 to 2.60, indicating that males are expected to be more likely to support AfD with the odds being 1.24 to 2.60. Education is also similar to the main model, with higher education being a significant predictor for AfD support in linear (CI: 1.21 to 2.31) and moderate predictor in quadratic (i.e., this effect accelarates as a decreasing rate) (CI: 0.35 to 0.92) case. Perceptions on immigrants' impact on economy also display a significant negative effect, with the confidence interval between 0.20 and 0.39, highlighting that as people's perception on immigration increases (getting positive), this is more likely to decrease their support for AfD. Additionally, the missingness indicators for variables like education and unemployment suggest that missing data can affect the relationship, with some indicators showing wide confidence intervals which do not provide a meaningful relationship with AfD support (e.g., imbgeco_ group_ missingness CI: 0.74 to 7.49).

10 Demonstration of the Performance of Statistical Tests

Since I am testing one hypothesis with different models (the main model with five covariates, the bivariate model, and the model with missingness), I first calculate the False Positive Rates (FPR) and then apply the Benjamini-Hochberg (BH) procedure. The results of the FPR show that approximately 77.8% of the predictors in the main model and 61.5% of the predictors in the model with missingness are expected to meet the criteria for inclusion at a 5% threshold[4]. In the bivariate model, which only includes one predictor, the FPR is 1, which makes it an unreliable measure because it only reflects whether the single predictor meets the threshold. The relatively high FPR values for the main model and the model with missingness suggest that there may be concerns regarding the inclusion of too many variables, which could lead to overfitting.

Given that FPR is limited in addressing the impact of overfitting and comparing multiple models, I move forward with the Benjamini-Hochberg (BH) procedure. This method helps control the False Discovery Rate (FDR) and false positive rates, ensuring more reliable conclusions about the relationships between variables. After applying this procedure, unemployment and gender emerge as the most important predictors for AfD support in the main model, with education,

[4] FPR values for main model, bivariate model, and model with missingness, respectively are: FPR= 0.7777777777777778; FPR=1; FPR=0.615384615384615.

religion, and economic perceptions on immigration also playing a significant role. Age does not appear to be a strong predictor in the main model. In the bivariate model, unemployment is identified as a key predictor, but without the inclusion of other variables, the effects of contextual factors (such as gender) are not accounted for. The model with missingness indicators follows a similar pattern to the main model, where the results for unemployment and other covariates align closely. Adding missingness variables did not notably affect the outcomes, except for the religion covariate. By using FPR followed by the BH procedure, I adjust for potential false discoveries and ensure that the multi-model analysis reflects the most accurate conclusions. Ultimately, the main model and the model with missingness suggest that unemployment and gender are key factors influencing AfD support, with missingness indicators are expected to have a minimal impact on the findings.

11 The Selection of Statistical Estimators and Estimand

Just as how I used in hypothesis testing, I use logistic regression to estimate the impact of prolonged unemployment on AfD (Alternative for Germany) support as the statistical estimator, based on Rooduijn (2008) and his research involving determinants of populist party support in a cross-national analysis. The estimand of interest is the odds ratio, which quantifies the likelihood of supporting AfD for individuals who have experienced unemployment for more than three months compared to those who have not. I choose odds ratio to facilitate interpretation of binary variables (both outcome and treatment) in my analysis. By including covariates of interest in this research such as gender, education, and economic perceptions on immigration, I adjust for potential confounders to isolate the effect of unemployment on AfD support. The odds ratio provides a direct interpretation of how the odds of supporting AfD are influenced by the treatment (unemployment), adjusted for the other variables in the model. Thus, the target estimand, which is the odds ratio, gives a good measure of the relationship between the binary treatment and the binary outcome, enabling me to analyze how unemployment affect political preferences while accounting for the influence of other factors.

12 Judgement Criteria for Statistical Estimators

I look at Mean Squared Error (MSE) values and bias as the judgement criteria for statistical estimators. This process starts with simulation of data to compare estimations of my model and true values. Accordingly, I simulated data for each model to see true values, and then run logistic regression to estimate parameters. As the next step, I compare true values with estimated coefficients. To calculate MSE values, I calculated squared error, which is the difference between true value and my estimated parameter. The average value of all these differences in simulations gives MSE values. Bias is a component of MSE value alongside variance, since it refers to the difference between true values and estimated parameters. This process is repeated for all three models.

13 Performance of Estimators via MSE and Bias

For main model and model with missingness, data frames of simulations only incorporated particular variables. In the main model, these variables were unemployment, age group and AfD support. I deliberately chose treatment and outcome variables and added another variable to represent other covariates. I made this choice to prevent any complexity that may occur by adding all variables. In the model with missingness, I included five main covariate variables to compare why having limited number of variables versus all covariates may make a difference.

When evaluating the performance of estimators across the three models (bivariate, main model, and the model with missingness) both Mean Squared Error (MSE) and bias provide interesting results. The lowest bias (-0.663) is the bivariate model, which also has a slightly lower MSE (2.397). This underscores that my bivariate model provides relatively stable estimates for the treatment effect (uemp3m) even if it only has one explanatory variable. Nevertheless, there is another side of the coin, simplicity in variable numbers limits model's ability to adjust for confounding effects, making causal inference mechanism difficult to interpret. The main model with five covariates, provides a balanced case between two other models with an MSE of 2.430 and a bias of -0.804. I can interpret that having the respective five covariates improved my concerns related to confounding effects and understanding the impact of the treatment effect but also presented slightly more variation in the estimates compared to the bivariate model. Finally, the model with missingness indicators had the highest MSE (2.467) and bias (-0.852). This reflects the additional noise from the inclusion of missingness variables. While these indicators aimed to account for potential biases arising from missing data, it also seems that they introduced instability in the treatment effect estimates.

These results display a good instance of trade-off before choosing adjustment strategies, adding covariates or missingness indicators: simpler models like the bivariate one has less variation but may neglect critical confounding factors, while more complex models, like the main and missingness models, address confounding effect better but mostly with some precision costs. To deal with possible issues on precision, I excluded some variables from the simulations to focus on the primary treatment and covariates critical to the hypothesis as I also stated earlier. This strategy minimized any complexities related to simulation computation process, to make sure that the analysis focused on variables directly related to AfD support and unemployment.

14 Mock Analysis

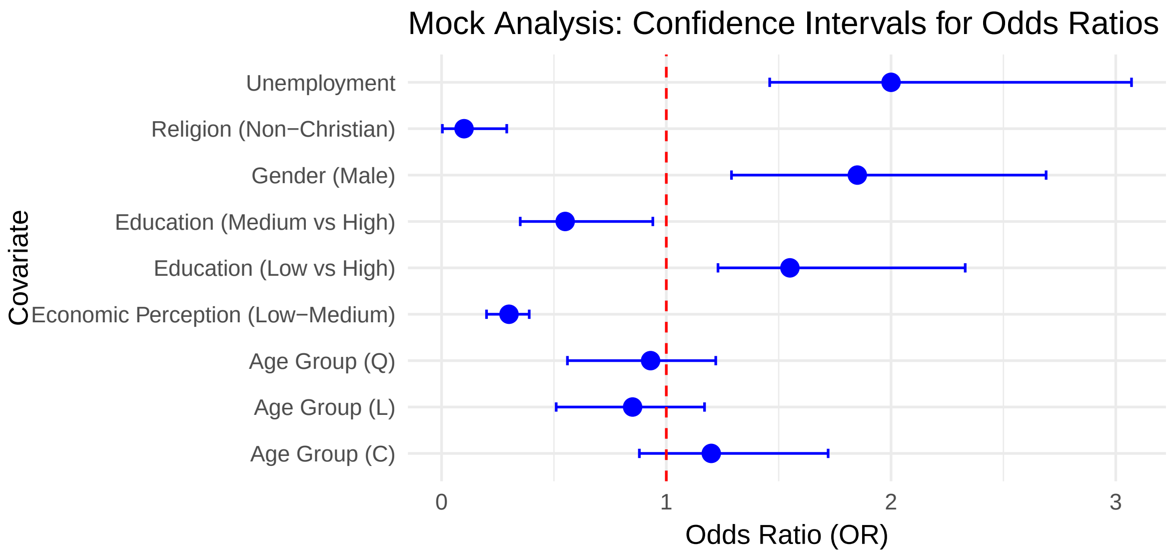

Figure 3: Confidence Intervals for Odds Ratios

If the real outcome mirrored these results, the figure would show both magnitude and direction of each covariate's impact on AfD support. The figure would also support my hypothesis that economic grievances, particularly within the context of prolonged unemployment, play a vital role in AfD support. Prolonged unemployment (OR = 2.00, 95% CI: 1.46–3.07) is expected to increase the likelihood of AfD support by 2 times compared to employment. This shows the weight of economic insecurities such as unemployment in party preferences. In addition, being male (OR = 1.85, 95% CI: 1.29–2.69) highlights that gender also positively associates with AfD support. Non-Christian respondents (OR = 0.10, 95% CI: 0.0035–0.29), however, are significantly less likely to support AfD. If I had these results in the real analysis, they would strongly support my hypothesis that prolonged unemployment plays a vital role in AfD support as the primary driver, but socio-cultural factors like gender and religion show a wider social landscape which is expected to shape support for right-wing populism in Germany.

15 Sensitivity Analysis

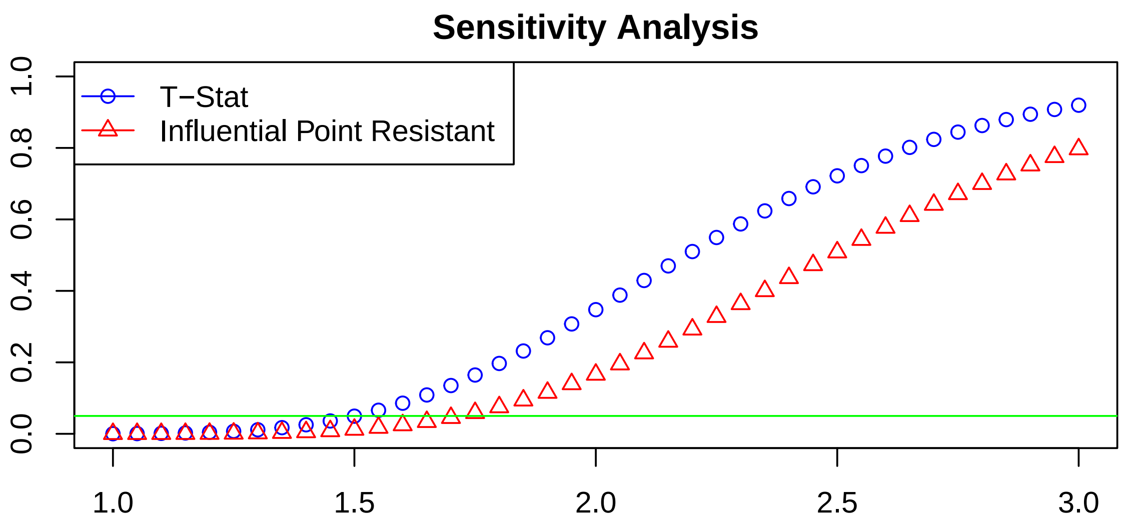

Figure 4: Sensitivity Analysis with Gamma Values

As the next step, I perform sensitivity analysis on main model with treatment and five covariates to see if there is an impact of unobserved covariates on the analysis. This is vital, because matching process only included five covariates apart from the main treatment variable, and there may be some other reasons impacting the relationship. I will do sensitivity analysis by following gamma as my sensitivity parameter based on Rosenbaum's (2005) contribution to the literature. I use senmv and reshape_ sensitivity functions in sensitivitymv package. The relevance of gamma parameter is that with the increasing values of gamma from 1 (no confounding) to higher numbers (e.g., 1.5, 2, 2.5, 3). Based on the change of gamma values, I will observe the change of p-value, which shows whether the relationship I see between unemployment and AfD support with five main covariates are due to random chance. If p-value increases as gamma values increase, this will display that the relationship is sensitive to unobserved counfounder effect.

The plot shows sensitivity analysis for two different methods: T-stat and Influential Point Resistant. The green line refers to the conventional significance threshold for p-value (0.05). As both methods show, when gamma values increase, p-values increase considerably (especially after gamma value of 1.5), highlighting that treatment effect gets less meaningful in the higher gamma values. This shows that treatment effect in this analysis might be sensitive to unobserved confounding with higher values of gamma.

16 Code Appendix

To see more details on the full code, please visit the following GitHub account: Saglam (2024).

```r
#EYLUL BEGUM SAGLAM

#QUANTITATIVE POLITICAL ANALYSIS II

#PRE-ANALYSIS PLAN

#DECEMBER 18, 2024

library(UsingR)

library(dplyr)

library(tidyverse)

library(ggplot2)

library(readxl)

library(car)

library(estimatr)

library(DeclareDesign)

library(MASS)

library(optmatch)

library(RItools)

library(robustbase)

library(highs)

library(designmatch)

library(coin)

library(haven)

library(readxl)

library(PanelMatch)

library(sensemakr)

library(sensitivitymv)
```

```
library(sensitivitymult)

library(sensitivityfull)

library(senstrat)

# Load and examine the dataset

ess11 <- read.csv("C:/Users/Eylul Begum Saglam/qpa/531-explorations-fall2024/

531-explorations-fall2024/ESS11-subset_2.csv")

# Check the structure and summary

str(ess11)

summary(ess11)

# Drop unnecessary variables

ess11_necessary_data <- ess11[, !(names(ess11) %in% c("psu", "stratum", "prob", "dweight",

"pweight", "prtvgde1", "prtclgde", "ctzcntr"))]

# Verify the structure after dropping

str(ess11_necessary_data)


# STEP 1: RECODE VARIABLES

## Treatment: uemp3m

ess11_necessary_data$uemp3m <- ifelse(ess11_necessary_data$uemp3m == 1, 1,

                                      ifelse(ess11_necessary_data$uemp3m == 2, 0, NA))

table(ess11_necessary_data$uemp3m)

## Outcome: prtvgde2

ess11_necessary_data$prtvgde2 <- ifelse(ess11_necessary_data$prtvgde2 == 6, 1,

        ifelse(ess11_necessary_data$prtvgde2 %in% c(1, 2, 3, 4, 5, 7, 8, 9, 55), 0, NA))

table(ess11_necessary_data$prtvgde2)

## Gender: gndr

ess11_necessary_data$gndr <- ifelse(ess11_necessary_data$gndr == 1, 1,

                                    ifelse(ess11_necessary_data$gndr == 2, 0, NA))
```

```
table(ess11_necessary_data$gndr)

# Recode prtvgde2 to make sure missing values are properly handled

ess11_necessary_data$prtvgde2[ess11_necessary_data$prtvgde2 %in% c(66, 77, 88, 99)] <- NA

## Religion: rlgdnade

# Recode religion into Christian and Non-Christian categories

ess11_necessary_data$rlgdnade_group <- ifelse(ess11_necessary_data$rlgdnade %in% c(1, 2,

                     3, 4, 290, 291), "Christian",

       ifelse(ess11_necessary_data$rlgdnade %in% c(5, 6, 7, 8, 9), "Non-Christian", NA))

# Verify the result

table(ess11_necessary_data$rlgdnade_group)

## Immigration economy: imbgeco

# Recode 'imbgeco' into "Low-Medium" (0-5) and "Medium-High" (6-10)

ess11_necessary_data$imbgeco_group <- ifelse(ess11_necessary_data$imbgeco %in% c(0, 1, 2,

                    3, 4, 5), "Low-Medium",

            ifelse(ess11_necessary_data$imbgeco %in% c(6, 7, 8, 9, 10), "Medium-High", NA))

# Verify the result

table(ess11_necessary_data$imbgeco_group)

## Education: edulvlb

ess11_necessary_data$edulvlb_group <- ifelse(ess11_necessary_data$edulvlb %in% c(0, 113,

129, 212, 213, 221, 222, 223), "Low",

ifelse(ess11_necessary_data$edulvlb %in% c(229, 311, 312, 313, 321, 322, 323, 412, 413, 421,

422, 423), "Medium",

ifelse(ess11_necessary_data$edulvlb %in% c(510, 520, 610, 620, 710, 720, 800), "High", NA)))

table(ess11_necessary_data$edulvlb_group)

# Categorize age into groups

ess11_necessary_data$age_group <- cut(ess11_necessary_data$agea,

                                breaks = c(0, 29, 49, 64, Inf),
```

```
                                labels = c("Young adults", "Adults", "Middle-aged", "Seniors"),
                                 right = TRUE) # right = TRUE includes the upper bound
# Check the distribution
table(ess11_necessary_data$age_group)
colSums(is.na(ess11_necessary_data))


# STEP 2: FACTORIZE SOME VARIABLES FOR ANALYSIS


ess11_necessary_data$uemp3m <- as.factor(ess11_necessary_data$uemp3m)
ess11_necessary_data$prtvgde2 <- as.factor(ess11_necessary_data$prtvgde2)
# Convert variables to factors
ess11_necessary_data$gndr <- as.factor(ess11_necessary_data$gndr)
ess11_necessary_data$rlgdnade_group <- as.factor(ess11_necessary_data$rlgdnade_group)
# Convert variables to ordered factors
ess11_necessary_data$edulvlb_group <- ordered(ess11_necessary_data$edulvlb_group)
ess11_necessary_data$imbgeco_group <- ordered(ess11_necessary_data$imbgeco_group)
ess11_necessary_data$age_group <- ordered(ess11_necessary_data$age_group)
colSums(is.na(ess11_necessary_data))


# STEP 3: MISSINGNESS EXPLORATION


library(VIM)
aggr_plot <- aggr(ess11_necessary_data, col = c('navyblue', 'red'),
                  numbers = TRUE, sortVars = TRUE, labels = names(ess11_necessary_data),
                  cex.axis = 0.7, gap = 3, ylab = c("Missing data", "Pattern"))
library(mice)
# Impute missing values using mice (multiple imputation)
```

```
# Step a: Specify the variables for imputation
imp_vars <- ess11_necessary_data[, c("rlgdnade_group", "imbgeco_group", "edulvlb_group",
            "uemp3m", "prtvgde2")]
# Step b: Perform multiple imputation using appropriate methods
imp_data <- mice(imp_vars, m = 5, method = c("logreg", "polr", "polr", "logreg", "logreg"),
            seed = 123)
# Step c: Create missingness indicators for imputed variables
ess11_necessary_data$uemp3m_missingness <- ifelse(is.na(ess11_necessary_data$uemp3m), 1, 0)
table(ess11_necessary_data$uemp3m_missingness)
ess11_necessary_data$prtvgde2_missingness <- ifelse(is.na(ess11_necessary_data$prtvgde2), 1,
0)
table(ess11_necessary_data$prtvgde2_missingness)
ess11_necessary_data$rlgdnade_group_missingness <- ifelse(is.na(ess11_necessary_data
$rlgdnade_group), 1, 0)
table(ess11_necessary_data$rlgdnade_group_missingness)
ess11_necessary_data$imbgeco_group_missingness <- ifelse(is.na(ess11_necessary_data$
imbgeco_group), 1, 0)
table(ess11_necessary_data$imbgeco_group_missingness)
ess11_necessary_data$edulvlb_group_missingness <- ifelse(is.na(ess11_necessary_data$
edulvlb_group), 1, 0)
table(ess11_necessary_data$edulvlb_group_missingness)
# Step d: Impute missing values for the main dataset
ess11_necessary_data$rlgdnade_group <- complete(imp_data, action = 1)$rlgdnade_group
ess11_necessary_data$imbgeco_group <- complete(imp_data, action = 1)$imbgeco_group
ess11_necessary_data$edulvlb_group <- complete(imp_data, action = 1)$edulvlb_group
ess11_necessary_data$prtvgde2 <- complete(imp_data, action = 1)$prtvgde2
# Impute prtvgde2
```

```r
ess11_necessary_data$uemp3m <- complete(imp_data, action = 1)$uemp3m

# Impute uemp3m if needed

# Step e: Check the summary of the imputed dataset

summary(ess11_necessary_data)

# Step f: Check for missing values again after imputation

colSums(is.na(ess11_necessary_data))

# Check the distribution of categorical variables after imputation

table(ess11_necessary_data$prtvgde2)

table(ess11_necessary_data$gndr)

table(ess11_necessary_data$rlgdnade_group)

table(ess11_necessary_data$imbgeco_group)

table(ess11_necessary_data$edulvlb_group)

table(ess11_necessary_data$age_group)

table(ess11_necessary_data$uemp3m)


# STEP 4: MATCHING


# Convert treatment variable to numeric (if necessary)

ess11_necessary_data$uemp3m <- as.numeric(as.character(ess11_necessary_data$uemp3m))

# Ensure all covariates are formatted appropriately

str(ess11_necessary_data)

# Define an exact matching constraint for one or more variables

# Example: Exact matching within levels of `gndr` and `rlgdnade_group`

exact_constraint <- exactMatch(uemp3m ~ gndr + rlgdnade_group, data = ess11_necessary_data)

# Compute propensity scores using a logistic regression model

ps_mod <- glm(uemp3m ~ imbgeco_group + edulvlb_group + age_group,

              data = ess11_necessary_data, family = binomial)
```

```
# Add the propensity scores to the dataset

ess11_necessary_data$ps <- predict(ps_mod, type = "response")

# Define the distance matrix for matching with a caliper on the propensity score

# The caliper restricts the matching distance to 0.2 of the standard deviation of the

propensity score

combined_distance <- match_on(ps_mod, data = ess11_necessary_data) +

  exact_constraint +  # Ensure exact matching within the constraints

  caliper(match_on(ps_mod, data = ess11_necessary_data), width = 0.2)  # Example caliper

  width of 0.2

# Use `fullmatch` to create matched groups

fm <- fullmatch(combined_distance, data = ess11_necessary_data)

#Add the matching results to the dataset

ess11_necessary_data$matched_set <- fm

#Summary of the matching

summary(fm)

#Tabulate treatment and control distribution within matched sets

table(ess11_necessary_data$uemp3m, ess11_necessary_data$matched_set, useNA = "ifany")

#Tabulate the number of treatment (1) and control (0) units in each matched set

table(ess11_necessary_data$uemp3m, ess11_necessary_data$matched_set, useNA = "ifany")

#Inspect the number of units in each matched set

table(ess11_necessary_data$matched_set)


# STEP 5: OMNIBUS TEST


#Hansen and Bowers

library(optmatch)

# Correct usage of xBalance
```

```r
balance_test <- xBalance(
  uemp3m ~ age_group + gndr + edulvlb_group + imbgeco_group + rlgdnade_group,
  strata = matched_data$matched_set,  # Include matched strata
  data = matched_data
)
# Print the results of the balance test
print(balance_test)
# Load necessary library
library(ggplot2)
# Create a dataframe with your balance test results
balance_results <- data.frame(
  Variable = c("age_groupYoung adults", "age_groupAdults", "age_groupMiddle-aged",
               "age_groupSeniors", "gndr0", "gndr1", "edulvlb_groupHigh",
               "edulvlb_groupLow", "edulvlb_groupMedium", "imbgeco_groupLow-Medium",
               "imbgeco_groupMedium-High", "rlgdnade_groupChristian",
               "rlgdnade_groupNon-Christian"),
  SMD = c(0.00, 0.01, 0.00, -0.01, 0.00, 0.00, 0.01, -0.03, 0.00, 0.02, -0.02, 0.00, 0.00),
  Z = c(-1.00, 2.03, -0.15, -1.29, 0.00, 0.00, 2.20, -2.14, 0.53, 2.50, -2.50, 0.00, 0.00)
)
# Add significance indicator based on Z-scores
balance_results$Significant <- ifelse(abs(balance_results$Z) > 1.96, "Significant",
"Not Significant")
# Plot standardized mean differences
ggplot(balance_results, aes(x = reorder(Variable, SMD), y = SMD, fill = Significant)) +
  geom_bar(stat = "identity", width = 0.6) +
  coord_flip() +  # Flip the axes for better readability
  scale_fill_manual(values = c("Not Significant" = "lightblue", "Significant" = "red")) +
```

```r
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  labs(
    title = "Standardized Mean Differences (SMD) After Matching",
    x = "Covariates",
    y = "Standardized Mean Difference (SMD)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1)) +
  guides(fill = guide_legend(title = "Significance"))
# Assuming `matched_data` is your dataset after matching
if (exists("matched_data")) {
  cat("Number of observations in the matched dataset:", nrow(matched_data), "\n")
} else {
  cat("Matched dataset not found. Please verify the variable name.\n")
}
# Check summary of matched data
summary(matched_data)


# STEP 6: POWER ANALYSIS (POST-MATCHING)


#Main Model:
# Define power analysis parameters
set.seed(123)
n_sim <- 1000          # Number of simulations
n_obs <- 719           # Effective sample size from matching
control_prob <- 0.2    # Baseline probability of AfD support in control group
treatment_effect <- 0.1  # Hypothesized increase in probability due to treatment
```

```r
alpha <- 0.05          # Significance level

# Placeholder to store p-values

p_values <- numeric(n_sim)

# Run simulations

for (i in 1:n_sim) {

  # Simulate covariates

  age_group <- factor(sample(c("Young", "Adult", "Middle-aged", "Seniors"), n_obs,

  replace = TRUE))

  gndr <- factor(sample(c("Male", "Female"), n_obs, replace = TRUE))

  edulvlb_group <- factor(sample(c("High", "Medium", "Low"), n_obs, replace = TRUE))

  imbgeco_group <- factor(sample(c("Low-Medium", "Medium-High"), n_obs, replace = TRUE))

  # Simulate treatment assignment (uemp3m)

  uemp3m <- rbinom(n_obs, 1, 0.5)  # 50% treatment probability

  # Simulate probabilities based on logistic regression

  logits <- log(control_prob / (1 - control_prob)) +

    uemp3m * log((control_prob + treatment_effect) / (1 - (control_prob + treatment_effect))

    )

  probabilities <- exp(logits) / (1 + exp(logits))

  # Generate binary outcomes

  prtvgde2 <- rbinom(n_obs, 1, probabilities)

  # Fit logistic regression

  model <- glm(prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group + imbgeco_group,

  family = binomial)

  # Extract p-value for treatment effect

  p_values[i] <- summary(model)$coefficients["uemp3m", 4]

}

# Calculate power: Proportion of p-values below significance threshold
```

```
power <- mean(p_values < alpha)

print(paste("Estimated Power:", power))

#Bivariate Model:

# Define parameters for the bivariate model

set.seed(123)

n_sim <- 1000          # Number of simulations

n_obs <- 719           # Effective sample size from matching

control_prob <- 0.2   # Baseline probability of AfD support in control group

treatment_effect <- 0.1  # Hypothesized increase in probability due to treatment

alpha <- 0.05          # Significance level

# Placeholder to store p-values

p_values_simple <- numeric(n_sim)

# Run simulations for simple model

for (i in 1:n_sim) {

  # Simulate treatment assignment (uemp3m)

  uemp3m <- rbinom(n_obs, 1, 0.5)  # 50% treatment probability

  # Simulate probabilities based on logistic regression

  logits <- log(control_prob / (1 - control_prob)) +

    uemp3m * log((control_prob + treatment_effect) / (1 - (control_prob + treatment_effect))

    )

  probabilities <- exp(logits) / (1 + exp(logits))

  # Generate binary outcomes

  prtvgde2 <- rbinom(n_obs, 1, probabilities)

  # Fit logistic regression (simple model)

  model_simple <- glm(prtvgde2 ~ uemp3m, family = binomial)

  # Extract p-value for treatment effect

  p_values_simple[i] <- summary(model_simple)$coefficients["uemp3m", 4]
```

```
}

# Calculate power for simple model: Proportion of p-values below significance threshold

power_simple <- mean(p_values_simple < alpha)

print(paste("Estimated Power for Simple Model:", power_simple))

#Missingness Model

# Define parameters for model with missingness

# Define parameters for model with missingness

set.seed(123)

n_sim <- 1000          # Number of simulations

n_obs <- 719           # Effective sample size from matching

control_prob <- 0.2   # Baseline probability of AfD support in control group

treatment_effect <- 0.1  # Hypothesized increase in probability due to treatment

alpha <- 0.05          # Significance level

# Placeholder to store p-values for model with missingness

p_values_missingness <- numeric(n_sim)

# Run simulations for model with missingness

for (i in 1:n_sim) {

  # Simulate treatment assignment (uemp3m)

  uemp3m <- rbinom(n_obs, 1, 0.5)  # 50% treatment probability

  # Simulate probabilities based on logistic regression

  logits <- log(control_prob / (1 - control_prob)) +

    uemp3m * log((control_prob + treatment_effect) / (1 - (control_prob + treatment_effect))

    )

  probabilities <- exp(logits) / (1 + exp(logits))

  # Generate binary outcomes

  prtvgde2 <- rbinom(n_obs, 1, probabilities)

  # Simulate covariates
```

```
  age_group <- factor(sample(c("Young", "Adult", "Middle-aged", "Seniors"), n_obs, replace =

  gndr <- factor(sample(c("Male", "Female"), n_obs,

  replace = TRUE))

  edulvlb_group <- factor(sample(c("High", "Medium", "Low"), n_obs, replace = TRUE))

  imbgeco_group <- factor(sample(c("Low-Medium", "Medium-High"), n_obs, replace = TRUE))

  # Simulate religion variable

  rlgdnade_group <- factor(sample(c("Christian", "Non-Christian"), n_obs, replace = TRUE))

  # Simulate missingness indicators

  imbgeco_group_missingness <- rbinom(n_obs, 1, 0.1)  # 10% missingness

  edulvlb_group_missingness <- rbinom(n_obs, 1, 0.1)

  uemp3m_missingness <- rbinom(n_obs, 1, 0.05)

  rlgdnade_group_missingness <- rbinom(n_obs, 1, 0.05)

  # Fit logistic regression (model with missingness)

  model_missingness <- glm(prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group +

                           rlgdnade_group + imbgeco_group + imbgeco_group_missingness +

                           edulvlb_group_missingness + uemp3m_missingness +

                           rlgdnade_group_missingness,

                         family = binomial)

  # Extract p-value for treatment effect (uemp3m)

  p_values_missingness[i] <- summary(model_missingness)$coefficients["uemp3m", 4]

}

# Calculate power for model with missingness: Proportion of p-values below significance

threshold

power_missingness <- mean(p_values_missingness < alpha)

print(paste("Estimated Power for Model with Missingness:", power_missingness))


# STEP 7: HYPOTHESIS TESTING
```

```r
#Main Model

# Logistic regression on matched data

model_logit <- glm(prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group + imbgeco_group +

                rlgdnade_group,

                   data = matched_data,

                   family = binomial)

summary(model_logit)

# Extract the coefficients and calculate the confidence intervals

conf_int <- confint(model_logit, level = 0.95)

# Exponentiate the coefficients to get odds ratios and their confidence intervals

odds_ratios <- exp(coef(model_logit))

odds_ratios_conf_int <- exp(conf_int)

# Print the odds ratios and their confidence intervals

print("Odds Ratios:")

print(odds_ratios)

print("Confidence Intervals for Odds Ratios:")

print(odds_ratios_conf_int)

#Bivariate

# Simple logistic regression model with only uemp3m (treatment) as predictor

simple_model <- glm(prtvgde2 ~ uemp3m, family = binomial, data = matched_data)

# Summary of the model

summary(simple_model)

# AIC of the simple model

simple_model_aic <- AIC(simple_model)

print(paste("AIC of the simple model: ", simple_model_aic))

conf_int <- confint(simple_model, level = 0.95)
```

```r
# Exponentiate the coefficients to get odds ratios and their confidence intervals

odds_ratios <- exp(coef(simple_model))

odds_ratios_conf_int <- exp(conf_int)

# Print the odds ratios and their confidence intervals

print("Odds Ratios:")

print(odds_ratios)

print("Confidence Intervals for Odds Ratios:")

print(odds_ratios_conf_int)

#Missingness

# Logistic regression model including missingness indicators for all covariates

model_with_missingness <- glm(prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group +

                                rlgdnade_group +

                                imbgeco_group + imbgeco_group_missingness +

                                edulvlb_group_missingness +

                                uemp3m_missingness + rlgdnade_group_missingness,

                              data = matched_data,

                              family = binomial)

# Summary of the regression model with missingness indicators

summary(model_with_missingness)

conf_int <- confint(model_with_missingness, level = 0.95)

# Exponentiate the coefficients to get odds ratios and their confidence intervals

odds_ratios <- exp(coef(model_with_missingness))

odds_ratios_conf_int <- exp(conf_int)

# Print the odds ratios and their confidence intervals

print("Odds Ratios:")

print(odds_ratios)

print("Confidence Intervals for Odds Ratios:")
```

```r
print(odds_ratios_conf_int)


# STEP 8: PERFORMANCE OF TESTS:


#False Positive Rates:

#Main Model:

# Extract p-values from your model

p_values <- summary(model_logit_2)$coefficients[, 4]  # Adjusted p-values (last column)

# Define significance threshold

threshold <- 0.05

# Count how many predictors have p-values below the threshold (significant predictors)

significant_predictors <- sum(p_values < threshold)

# Calculate total number of predictors (excluding the intercept)

total_predictors <- length(p_values) - 1

# Calculate false positive rate (FPR)

FPR <- significant_predictors / total_predictors

# Print the result

print(paste("False Positive Rate (FPR):", FPR))

#Bivariate Model:

# Bivariate model FPR for single predictor

p_values_2 <- summary(simple_model_2)$coefficients[, 4]  # Adjusted p-values (last column)

# Define significance threshold

threshold <- 0.05

# Check if the only predictor (uemp3m) is significant

FPR <- ifelse(p_values_2[2] < threshold, 1, 0)

# Only check uemp3m, since it's the only predictor apart from the intercept

# Print the result
```

```r
print(paste("False Positive Rate (FPR):", FPR))

#Model with Missingness:

p_values_3 <- summary(model_with_missingness_2)$coefficients[, 4]

# Adjusted p-values (last column)

# Define significance threshold

threshold <- 0.05

# Count how many predictors have p-values below the threshold (significant predictors)

significant_predictors <- sum(p_values_3 < threshold)

# Calculate total number of predictors (excluding the intercept)

total_predictors <- length(p_values_3) - 1

# Calculate false positive rate (FPR)

FPR <- significant_predictors / total_predictors

# Print the result

print(paste("False Positive Rate (FPR):", FPR))

#BH

#Main Model

#Fit your logistic regression model

model_logit_2 <- glm(prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group +

               imbgeco_group + rlgdnade_group, family = binomial, data = matched_data)

# Get the summary of the model

summary_model <- summary(model_logit_2)

# Extract the p-values from the model summary

p_values <- summary_model$coefficients[, 4]  # 4th column contains the p-values

# Adjust the p-values using the Benjamini-Hochberg procedure (FDR control)

p_adjusted <- p.adjust(p_values, method = "BH")

# Print the adjusted p-values

print(p_adjusted)
```

```r
#Bivariate model

simple_model_2 <- glm(prtvgde2 ~ uemp3m, family = binomial, data = matched_data)

# Get the summary of the model

summary_model_2 <- summary(simple_model_2)

# Extract the p-values from the model summary

p_values_2 <- summary_model_2$coefficients[, 4]  # 4th column contains the p-values

# Adjust the p-values using the Benjamini-Hochberg procedure (FDR control)

p_adjusted_2 <- p.adjust(p_values_2, method = "BH")

# Print the adjusted p-values

print(p_adjusted_2)

#Model with Missingness

model_with_missingness_2 <- glm(prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group +

                                  rlgdnade_group +

                                  imbgeco_group + imbgeco_group_missingness +

                                  edulvlb_group_missingness +

                                  uemp3m_missingness + rlgdnade_group_missingness,

                                data = matched_data,

                                family = binomial)

summary_model_3 <- summary(model_with_missingness_2)

# Extract the p-values from the model summary

p_values_3 <- summary_model_3$coefficients[, 4]  # 4th column contains the p-values

# Adjust the p-values using the Benjamini-Hochberg procedure (FDR control)

p_adjusted_3 <- p.adjust(p_values_3, method = "BH")

# Print the adjusted p-values

print(p_adjusted_3)


# STEP 9: MSE and Bias
```

```r
#Main Model:

# Define parameters

n_sim <- 1000

mse_values <- numeric(n_sim)  # Store MSE for each simulation

# True value of the parameter

true_param <- 1.5

# Simulate data and calculate MSE

for (i in 1:n_sim) {

  # Simulate data

  simulated_data <- data.frame(

    uemp3m = rbinom(100, 1, 0.5),

    age_group = sample(c("Young", "Adult", "Middle-aged", "Seniors"), 100, replace = TRUE),

    prtvgde2 = rbinom(100, 1, 0.5)

  )

  # Fit logistic regression model

  model <- glm(prtvgde2 ~ uemp3m + age_group, data = simulated_data, family = binomial)

  # Estimate parameter for 'uemp3m'

  estimated_param <- coef(model)["uemp3m"]

  # Calculate MSE for this simulation

  mse_values[i] <- (estimated_param - true_param)^2  # MSE is just the squared error here

}

# Calculate average MSE

mean_mse <- mean(mse_values)

print(paste("Mean Squared Error (MSE):", mean_mse))

# Simulating true parameters

true_param <- 1.5  # True value of the coefficient you're estimating
```

```
# Run logistic regression model on simulated data

model <- glm(prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group + imbgeco_group,

             data = matched_data, family = binomial)

# Estimate parameter

estimated_param <- coef(model)["uemp3m"]  # Extract the estimated coefficient for the treatme

# Calculate bias

bias <- estimated_param - true_param

print(paste("Bias: ", bias))

#Bivariate

# Define parameters for bivariate model

n_sim <- 1000

mse_values_bivariate <- numeric(n_sim)  # Store MSE for each simulation

# True value of the parameter for bivariate model

true_param_bivariate <- 1.5

# Simulate data and calculate MSE for bivariate model

for (i in 1:n_sim) {

  # Simulate data

  simulated_data <- data.frame(

    uemp3m = rbinom(100, 1, 0.5),  # Simulate treatment

    prtvgde2 = rbinom(100, 1, 0.5) # Simulate binary outcome

  )

  # Fit logistic regression model

  model_bivariate <- glm(prtvgde2 ~ uemp3m, data = simulated_data, family = binomial)

  # Estimate parameter for 'uemp3m'

  estimated_param_bivariate <- coef(model_bivariate)["uemp3m"]

  # Calculate MSE for this simulation

  mse_values_bivariate[i] <- (estimated_param_bivariate - true_param_bivariate)^2
```

```
}

# Calculate average MSE for bivariate model

mean_mse_bivariate <- mean(mse_values_bivariate)

print(paste("Mean Squared Error (MSE) for Bivariate Model:", mean_mse_bivariate))

# Calculate bias for bivariate model

model_bivariate_real <- glm(prtvgde2 ~ uemp3m, data = matched_data, family = binomial)

estimated_param_bivariate_real <- coef(model_bivariate_real)["uemp3m"]

bias_bivariate <- estimated_param_bivariate_real - true_param_bivariate

print(paste("Bias for Bivariate Model:", bias_bivariate))

#Missingness

# Define parameters for model with missingness

n_sim <- 1000

mse_values_missingness <- numeric(n_sim)  # Store MSE for each simulation

# True value of the parameter for model with missingness

true_param_missingness <- 1.5

# Simulate data and calculate MSE for model with missingness

for (i in 1:n_sim) {

  # Simulate data

  simulated_data <- data.frame(

    uemp3m = rbinom(100, 1, 0.5),  # Simulate treatment

    age_group = sample(c("Young", "Adult", "Middle-aged", "Seniors"), 100, replace = TRUE),

    gndr = sample(c("Male", "Female"), 100, replace = TRUE),  # Simulate gender

    edulvlb_group = sample(c("High", "Medium", "Low"), 100, replace = TRUE),

    # Simulate education level

    imbgeco_group = sample(c("Low-Medium", "Medium-High"), 100, replace = TRUE),

    # Simulate economic perception

    prtvgde2 = rbinom(100, 1, 0.5) # Simulate binary outcome
```

```
  )

  # Fit logistic regression model with missingness

  model_missingness <- glm(

    prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group + imbgeco_group,

    data = simulated_data,

    family = binomial

  )

  # Estimate parameter for 'uemp3m'

  estimated_param_missingness <- coef(model_missingness)["uemp3m"]

  # Calculate MSE for this simulation

  mse_values_missingness[i] <- (estimated_param_missingness - true_param_missingness)^2

}

# Calculate average MSE for model with missingness

mean_mse_missingness <- mean(mse_values_missingness)

print(paste("Mean Squared Error (MSE) for Model with Missingness:", mean_mse_missingness))

# Calculate bias for model with missingness

model_missingness_real <- glm(

  prtvgde2 ~ uemp3m + age_group + gndr + edulvlb_group + imbgeco_group +

    imbgeco_group_missingness + edulvlb_group_missingness + uemp3m_missingness

    + rlgdnade_group_missingness,

  data = matched_data,

  family = binomial

)

estimated_param_missingness_real <- coef(model_missingness_real)["uemp3m"]

bias_missingness <- estimated_param_missingness_real - true_param_missingness

print(paste("Bias for Model with Missingness:", bias_missingness))
```

```r
# STEP 10: MOCK ANALYSIS

# Load necessary library
# Mock data: Odds Ratios and Confidence Intervals, including Age Groups
# Load necessary library
library(ggplot2)
# Data for covariates and their confidence intervals (from your results)
mock_data <- data.frame(
  Covariate = c("Intercept", "Unemployment", "Age Group (L)", "Age Group (Q)",
                "Age Group (C)", "Gender (Male)", "Education (Low vs High)",
                "Education (Medium vs High)", "Economic Perception (Low-Medium)",
                "Religion (Non-Christian)"),
  OR = c(NA, 2.00, 0.85, 0.93, 1.20, 1.85, 1.55, 0.55, 0.30, 0.10),
  # Odds Ratios from CI midpoints
  Lower_CI = c(0.02, 1.46, 0.51, 0.56, 0.88, 1.29, 1.23, 0.35, 0.20, 0.0035),
  # Lower bounds
  Upper_CI = c(0.04, 3.07, 1.17, 1.22, 1.72, 2.69, 2.33, 0.94, 0.39, 0.29)
  # Upper bounds
)
# Exclude the intercept for visualization (optional)
mock_data_no_intercept <- mock_data[-1, ]
# Create the visualization
ggplot(mock_data_no_intercept, aes(x = Covariate, y = OR)) +
  geom_point(size = 3, color = "blue") +
  geom_errorbar(aes(ymin = Lower_CI, ymax = Upper_CI), width = 0.2, color = "blue") +
  geom_hline(yintercept = 1, linetype = "dashed", color = "red") +
```

```
  theme_minimal() +

  coord_flip() +  # Flip coordinates for better readability

  labs(

    title = "Mock Analysis: Confidence Intervals for Odds Ratios",

    x = "Covariate",

    y = "Odds Ratio (OR)"

  )


# STEP 11: SENSITIVITY ANALYSIS


# Sensitivity analysis using different gamma values

library(sensemakr)

library(sensitivitymv)

library(sensitivitymw)

library(sensitivitymult)

library(sensitivityfull)

library(senstrat)

library(rbounds)

# Gamma values to test

somegammas <- seq(1, 3, 0.05)

# Reshape the data for sensitivity analysis

respmat <- with(matched_data, reshape_sensitivity(

  y = prtvgde2,                # Outcome variable (AfD support)

  z = uemp3m,                  # Treatment variable (Unemployed vs. Employed)

  fm = matched_set             # Matching strata from full matching

))

# Perform sensitivity analysis with t-statistic (method="t")
```

```
sensTresults <- sapply(somegammas, function(g) {
  c(gamma = g, senmv(respmat, method = "t", gamma = g)$pval)
})
# Perform sensitivity analysis with mean difference (method="mean")
sensHresults <- sapply(somegammas, function(g) {
  c(gamma = g, senmv(respmat, gamma = g)$pval)
})
print(sensHresults)
# Adjusting plot margins and layout
# Check the structure of sensTresults and sensHresults
# Extract gamma values and p-values for sensTresults
gamma_T <- sensTresults[1, ]  # First row: gamma values
pval_T <- sensTresults[2, ]   # Second row: p-values
# Extract gamma values and p-values for sensHresults
gamma_H <- sensHresults[1, ]  # First row: gamma values
pval_H <- sensHresults[2, ]   # Second row: p-values
# Adjusting plot margins and layout
par(mar = c(3, 3, 2, 1), mfrow = c(1, 1))
# Plot Sensitivity Analysis Results (T-stat)
plot(
  x = gamma_T,
  y = pval_T,
  type = "b",  # type 'b' for both points and lines
  xlab = "Gamma",
  ylab = "P-Value",
  main = "Sensitivity Analysis",
  col = "blue",  # color for the first plot (sensTresults)
```

```
  pch = 1,  # point type for the first plot

  lty = 1,  # line type for the first plot

  ylim = c(0, 1)  # y-axis range (p-values range from 0 to 1)

)

# Add second set of points (Influential Point Resistant) to the same plot

points(

  x = gamma_H,

  y = pval_H,

  pch = 2,  # different point type for the second set

  col = "red"  # color for the second plot (sensHresults)

)

# Add a horizontal red line at p-value = 0.05 (significance threshold)

abline(h = 0.05, col = "green")  # Dashed red line at 0.05

# I add legends or other customizations to enhance the plot.

legend("topleft", legend = c("T-Stat", "Influential Point Resistant"),

       col = c("blue", "red"), pch = c(1, 2), lty = 1)
```

References

Bernburg, Jón Gunnar. "Economic crisis and popular protest in Iceland, January 2009: The role of perceived economic loss and political attitudes in protest participation and support". *Mobilization: An International Quarterly* 20, no. 2 (2015): 231–252.

Bowers, Jake. "Final Project Template for PS 531: A Pre-Analysis Plan". *GitHub* (2024). Visited on 12/12/2024. `https://github.com/bowers-grad-stats-illinois/531-syllabus/blob/main/final_paper_template.md`.

Chen, Lula, and Chris Grady. "10 Things to Know About Pre-Analysis Plans". *EGAP*, visited on 12/12/2024. `https://egap.org/resource/10-things-to-know-about-pre-analysis-plans/`.

Clarke, Seán. "German election opinion polls – who's leading for 2025". *The Guardian* (Dec. 12, 2024). Visited on 12/12/2024. `https://www.theguardian.com/world/ng-interactive/2024/dec/12/german-election-opinion-polls-whos-leading-for-2025`.

European Commission. "International Standard Classification of Education (ISCED)". Visited on 10/10/2024. `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCED)#:~:text=ISCED%20is%20the%20reference%20international,has%20been%20implemented%20since%202016.`.

European Social Survey (ESS). "ESS round 11 - 2023. Social inequalities in health, Gender in contemporary Europe", 2024. Visited on 11/30/2024. `https://ess.sikt.no/en/study/412db4fe-c77a-4e98-8ea4-6c19007f551b/132`.

Federal Ministry of the Interior and Community. "The voting system". Visited on 11/30/2024. `https://www.bmi.bund.de/EN/topics/constitution/electoral-law/voting-system/voting-system-node.html`.

Golder, Matt. "Far right parties in Europe". *Annual review of political science* 19, no. 1 (2016): 477–497.

Grasso, Maria T, and Marco Giugni. "Protest participation and economic crisis: The conditioning role of political opportunities". *European Journal of Political Research* 55, no. 4 (2016): 663–680.

Hansen, Ben B, and Jake Bowers. "Covariate balance in simple, stratified and clustered comparative studies". *Statistical Science* (2008): 219–236.

Kern, Anna, Sofie Marien, and Marc Hooghe. "Economic crisis and levels of political participation in Europe (2002–2010): The role of resources and grievances". *West European Politics* 38, no. 3 (2015): 465–490.

Kurer, Thomas, et al. "Economic grievances and political protest". *European Journal of Political Research* 58, no. 3 (2019): 866–892.

Rabb, Nathaniel, et al. "The influence of social norms varies with "others" groups: Evidence from COVID-19 vaccination intentions". *Proceedings of the National Academy of Sciences* 119, no. 29 (2022): e2118770119.

Rooduijn, Matthijs. "What unites the voter bases of populist parties? Comparing the electorates of 15 populist parties". *European Political Science Review* 10, no. 3 (2018): 351–368.

Rosenbaum, Paul R. "Sensitivity analysis in observational studies". *Encyclopedia of statistics in behavioral science* 4 (2005): 1809–1814.

Rosenbaum, Paul R, P Rosenbaum, and Briskman. *Design of observational studies.* Vol. 10. Springer, 2010.

Rüdig, Wolfgang, and Georgios Karyotis. "Who protests in Greece? Mass opposition to austerity". *British Journal of Political Science* 44, no. 3 (2014): 487–513.

Saglam, Eylul Begum. "Pre-Analysis-Plan/FinalPaperCodes.R". *GitHub* (2024). Visited on 12/18/2024. `https://github.com/esaglam2/Pre-Analysis-Plan/blob/main/FinalPaperCodes.R`.

Schwander, Hanna, and Philip Manow. "It's not the economy, stupid! Explaining the electoral success of the German right-wing populist AfD". *CIS Working Paper*, no. 94 (2017).

Silver, Laura. "Populists in Europe – especially those on the right – have increased their vote shares in recent elections". *Pew Research Center* (June 10, 2022). Visited on 12/12/2024. `https://www.pewresearch.org/short-reads/2022/10/06/populists-in-europe-especially-those-on-the-right-have-increased-their-vote-shares-in-recent-elections/`.