Predicting EU-skepticism in France

STV4020A – Forskningsmetode og statistikk

Høst 2019

Kandidatnummer: 17521

Antall ord: 3968

Datasett: European Social Survey, runde 8

Introduction

In this paper I want to investigate whether or not economic confidence and satisfaction can predict support for the European Union. The research question is based on cost-benefit theory, which is widely applied in the study of European integration. I will control for anti-immigration attitudes and relevant socio-economic factors. The research question is: *Are people who do not feel confident about their economic future and the country's economic future less likely to support EU-membership?* To answer the research question, I will use logistical regression. The results presented in this paper finds a correlation between evaluating one's economic outlook negatively and EU-skepticism.

Data from the French European Social Survey round 8 will be used to answer the research question. My main motivation for researching this topic is to investigate how important economic aspects are for determining support for the European Union. In light of the rise of the radical right party Rassemblement National and the newly founded radical left party (and also EU-skeptic) La France Insoumise, France serves as an interesting case for testing the theory.

The main focus in this paper is the method applied. The paper is organized as follows: First some preliminary theory to add some background. Following this, an overview of the dataset and the procedure from theory to measurable concept. Then I turn to the results – the answer of the research question. A thorough review of the method applied, diagnostics and goodness of fit tests comes naturally after presenting the results. The final section of the paper briefly discusses the findings in light of Cook and Campbell's validity system.

Theory

People who feel confident about their economic future – personally, and for their country, are more likely to regard European integration in a positive light (Hooghe & Marks, 2005). Whereas those who are fearful of their economic future and the country's economic future are more EU-skeptic (ibid.). This means that how individuals view personal economic prospects and the country's economic prospects affect their attitudes towards European integration.

Hooghe & Marks (2005) separate the economic models of public opinion on European integration into two categories: objective and subjective evaluation. For the purpose of this paper I am focusing on subjective evaluation: people's feelings about their own personal and their country's economic situation. Thus, the attention of this paper will be directed towards

subjective economic evaluation on the egocentric and sociotropic level (personal and national economic prospects).

De Vreese & Boomgaarden (2005) show that anti-immigration attitudes are associated with EU-skepticism. Therefore, it is necessary to control for attitudes towards immigration, in order to better estimate the effect of economic outlooks.

Hypothesis

According to the cost-benefit theory people who feel more confident about their personal economic future of their country's economic future are more likely to support European integration. This means that people who feel less confident about their personal economic future and their country's economic future are less likely to support European integration. Whether or not this translates into wanting to leave the EU is what I am investigating in this paper. Therefore, the hypothesis' builds upon the assumptions from the cost-benefit theory:

> H0: *People who are not confident about their personal economic future and their country's economic future are not more likely to support leaving the European Union.*

> H1: *People who are not confident about their personal economic future and their country's economic future are more likely to support leaving the European Union.*

Data

European Social Survey is a cross-national survey conducted every two years with data from over 23 countries.[1] The survey is a product of collaboration between different universities and the survey measures attitudes, beliefs and behavior patterns. ESS has been conducted since 2001, but for the purpose of this paper I will be using round 8 and the country specific version for France. In the country specific dataset for France there are 2070 observations and 540 variables. The data from round 8 was collected between July 2016 to April 2017 (Wuyts & Loosveldt, 2019, p. 8).

Dependent variable

EU-skepticism can in broad terms be defined as opposition to European integration (Halikiopoulou et al, 2012). In order to measure attitudes towards the EU I will use a variable that asks respondents to answer if they would vote for their country to leave the EU or if they

[1] For more information about European Social Survey: https://www.europeansocialsurvey.org/about/faq.html

would vote for their country to remain a member of the EU. This variable is used to measure EU-skepticism because it is highly unlikely that a person that is not skeptical towards the EU would cast a vote to leave the EU or would on any level want to leave the EU. Thus, this variable is used in order to ensure strong concept validity. The variable is binary: vote to remain in the EU is coded 0 and vote to leave the EU is coded 1.

Independent variables

The cost benefit theory of European integration identifies positive subjective evaluation of economic outlook as a driver for positive attitudes towards the European Union (Hooghe & Marks, 2005). Feelings tied to economic future will be affected by a person's evaluation of the current situation and what that situation may lead to in the future. In order to be able to capture subjective evaluation of personal and country's economic outlook the following variables will be included in the model:

1) *How likely not enough money for household necessities next 12 months (money)* – scale: 0 = not very likely, 3 = very likely.
2) *How satisfied with the country's economy (economy)* – scale reversed, therefore in effect I am measuring dissatisfaction. 10 = not satisfied at all, 0 = satisfied.

Variable 2, *economy,* is more immediate as it is asking about satisfaction with the economy at the time of data collection. This is contrasting with the other variable which is asking about personal economic future in the next 12 moths. Nevertheless, this variable is included because individual's that will be dissatisfied with the current state of the economy, would not view the country's economic prospects as positive in the near future either.

In the analysis I will be controlling for socio-economic features, such as age, gender and years of education *(agea, gender, eduyrs).* Gender has been coded to man = 1, and women = 0. The education variable asks how many years of completed education the respondent has undertaken. *Agea* is calculated from respondents date of birth.

I will also control for anti-immigration attitudes. De Vreese & Boomgaarden (2005) argues that there is a connection between EU-skepticism and anti-immigration attitudes. Therefore, including a variable that measures immigration attitudes will allow for better estimation of the above mentioned economic explanatory variables. The immigration variable *(immigration1)* included in this model asks respondents if they think immigrants make the country a worse or better place to live on a scale from zero to ten, where ten is better place to

live and zero worse place to live. The scale has been reversed, where ten has been coded to worse place to live, indication that in effect I am measuring anti-immigration attitudes. The scales have been reversed to allow for meaningful interpretation of the results and to ensure that the explanatory variables have scales that points in the same direction.

Method of analysis

In order to be able to answer the research question the method of analysis will be logistical regression. This is because the dependent variable is binary (vote to leave the EU or not). Logistical regression models the probability that the dependent variable equals 1, *Pr(Y=1)*. I will use the logit-model and the logit-model coefficients are estimated using maximum likelihood. Maximum likelihood chooses the values of the parameters to maximize the probability of drawing the data that are actually observed (Stock & Watson, 2015, p. 446). In this sense, the maximum likelihood estimators are the parameter values "most likely" to have produced the data (Stock & Watson, 2015, p. 446).

Observations with missing on one or several of the variables included is removed from the analysis. The listwise deletion ensures that the results are consistent as long as the observations are missing completely at random (MCAR) (Christophersen, 2013, p. 82). Assuming the missing observations are MCAR and not missing at random (MAR) is because missing on one of the variables is not determined by missingness on another variable included in the analysis. Listwise deletion is selected because it is hard to conduct imputation on a binary variable and because the missing values are assumed to be MCAR.

The dataset is split into two parts in order to build my model using 75 percent of the dataset, whilst a smaller part of the dataset (the remaining 25 percent) is used to evaluate my model. Selecting observations for each dataset is done through random sampling. Splitting the dataset will allow me to evaluate how well my model predicts on new and unseen data. Testing predictive performance on out-of-data observations is an effective way to determine the model's external validity and reliability.

Results

The results are found in figure 1. I run two regressions, one model where the variable *immigration1* is included and one where it is not in order to see how this variable affects the estimation of the other variables. To begin with age and years of education is negatively correlated with for France to leave the EU. Whilst the other variables are positively correlated. In model 2 gender is not significant and in models 1 it is significant at the 5

percent level. All other estimated coefficients, except from age, are significant at the 1 percent level. Age is significant at the 5 percent level in both models. What is interesting to note from the table below is how my explanatory variables change when immigration attitudes are controlled for.

Regression Results

	Dependent variable:	
	Would vote for France to remain a member of European Union or leave	
	(1)	(2)
Age	-0.010**	-0.012**
	(0.005)	(0.005)
Man	0.303**	0.253*
	(0.144)	(0.151)
Years of education	-0.103***	-0.065***
	(0.022)	(0.023)
Satisfaction with economy	0.322***	0.248***
	(0.039)	(0.041)
Likely not enough money	0.343***	0.350***
	(0.084)	(0.088)
Anti-immigration attitudes		0.321***
		(0.037)
Constant	-2.130***	-3.722***
	(0.523)	(0.573)
Observations	1,247	1,247
Log Likelihood	-613.348	-570.451
Akaike Inf. Crit.	1,238.697	1,154.901
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 1.0: Results from multiple logistical regression. Estimated coefficients are in logit.

The results from model 2 shows that the odds of voting in favor of France leaving the EU increases with 28 percent if dissatisfaction with the economy increases with one unit, holding everything else constant.[2] The odds of voting in favor of France leaving the EU increases with 42 percent if likeliness of not having enough money for household necessities increases with one unit, ceteris paribus. The odds of voting in favor of France leaving the EU increases with 38 percent if anti-immigration attitudes increase with one unit, ceteris paribus. In model 1 the odds of voting leave increases with 38 percent (10 percent more than in model 2) if dissatisfaction with the economy increases with one unit, holding everything else equal. Whilst the odds of voting leave are 41 percent (actually 1 percent lower than model 2) if likeliness of for not having enough money for household necessities increases with one unit, ceteris paribus. The estimated variable *likely not enough money* increases from model 1 to

[2] Results reported here are rounded to nearest whole number.

model 2, when anti-immigration attitudes are controlled for. This could perhaps mean that there is an interaction between anti-immigration attitudes and likely not enough money. Indicating that people that score high on the anti-immigration variable tend to be more pessimistic about their personal economic future.

Conclusively, hypothesis 1 – *people who are not confident about their personal economic future and their country's economic future are more likely to support leaving the European Union* – in light of the findings, should not be rejected. However, as suspected, anti-immigration attitudes also explain a lot of EU-skepticism.

Diagnostics and model evaluation

The underlying assumptions of the logistical regression model are that the dependent variable is binary; the probability curve is S-shaped and the logit curve is linear; there are no influential observations; there is no multicollinearity among the predictors; there are no empty cells; there is no "complete separation"; no omitted variable bias and the observations are independent and identically distributed.

The first assumption holds as the dependent variable is binary: "Would vote for France to leave the EU" has been coded to 0 = *would vote to remain*, 1 = *would vote to leave.* The assumption that the observations are independent and identically distributed holds if the sample is gathered using random probability selection. ESS sample uses random probability methods at all stages (Wuyts & Loosveldt, 2019, p. 9). No omitted variable bias is an important theoretical assumption. Arguably the assumption is met in model 2 because this model includes an additional relevant control variable. Nevertheless, there are many other features associated with EU-skepticism which neither of my models controls for, like for example the importance of national identity and political affiliation.

In order to investigate if the second assumption holds, I have to make sure that the relationship between the independent variables and the logit outcome is linear. The figure below matches a logit curve to each individual variable in order to see if the logit curve is linear. Figure 2.0 shows that the linearity assumption holds.

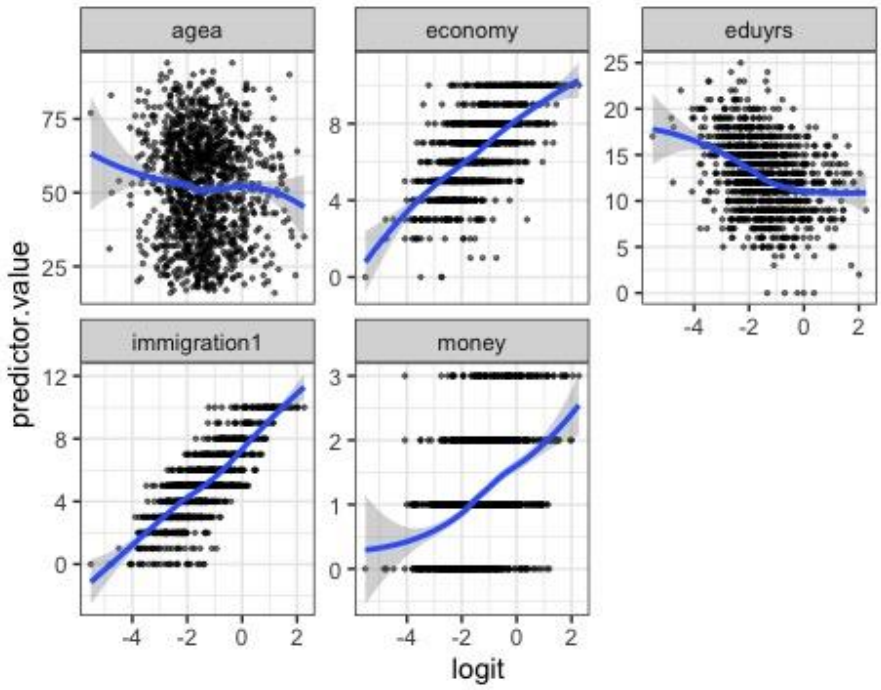


Figure 1.0 illustrates that the independent variables do have a linear relationship on the logit scale. The gender variable is not included because it is binary.

In order to check if the data contains any influential observations plotting the residuals is helpful. Standardizing the residuals allows for a simpler classification of influential observations. The standardized normal distribution lies between -4 and 4. Values above 3 indicates outliers and should be further investigated as they might affect the result.

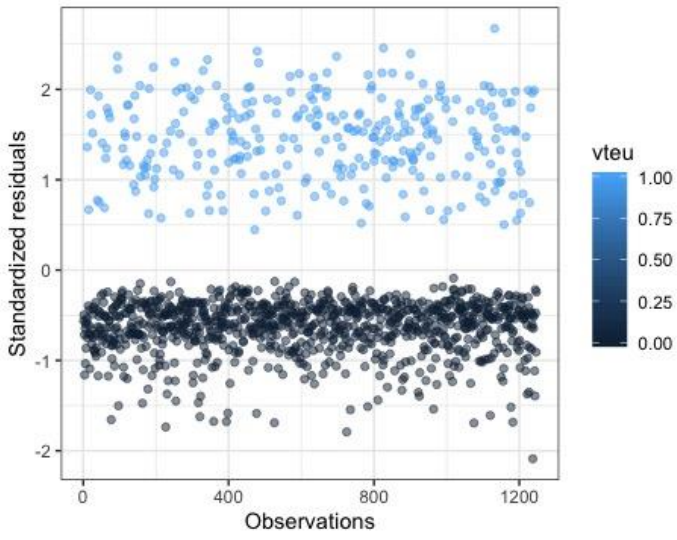


Figure 2.0 shows the standardized residuals for each observation.

The figure is a plot of the residual to all observations. No observations have a residual larger than 3 standard deviations from the mean, indicating no influential values.

The VIF-test is useful in detecting multicollinearity. VIF-test measures how much of the variance in each independent variable can be explained by the other variables in the analysis. General rule of thumb is a VIF-value under 5 indicates no multicollinearity, whilst values between 5 and 10 is considered not ideal, but not very problematic (Hermansen, 2019, p. 195). Values above 10 indicates strong multicollinearity (ibid.). Running the VIF-test I find that all my independent variables have VIF-values between 1 and 2, which indicates no multicollinearity.

Checking for complete separation is easily done by plotting the data. The plot indicates that this assumption also is met. The results from this procedure is found in the appendix. Additionally, checking for empty cells is unnecessary as observations with missing is not included in the model.

Goodness of fit

McFadden's pseudo R2 is a measure that compares the log-likelihood value for my model and compares it to the log-likelihood value for a model with no variables – an intercept-only-model (Christophersen, 2013, p. 139). The value ranges from zero to one. Values closer to 1 indicates good predictive power. Values close to zero indicates no predicative power. The McFadden's pseudo R2-value for model 2 is 0.173. The pseudo R2 for model 1, which is the model without *immigration1* as control variable, is: 0.11. Neither of the models explain a lot of the variance. The values are closer to 0 than to 1. Nevertheless, when immigration is included in the model the model is able to explain more of the variance.

As can be read from the results table presented above the loglikelihood value in model 2 is smaller than the loglikelihood value in model 1. This indicates that model 2 is better fitted to the data (Christophersen, 2013, p. 138). The difference between model 1 and model 2 is statistically significant according to the loglikelihood ratio test, showcasing that model 2 is better fitted to the data.

The Hosmer-Lemeshow-test tests how good the model fits the data by comparing observed and predicted values – meaning that it compares the observed, real values of 1 and 0, to the models fitted values (ibid.). The test does this by comparing subgroups of the population estimated. The Hosmer-Lemeshow-test is not supposed to give significant results, because this means that the model is not a good fit for the data. The p-values for model 1 and

[3] Value for McFadden's pseudo R^2 is rounded to nearest two decimals

2 is well above 0.05, which means that the results are not significant, and the model is good at describing the data.

How well does my model predict?

In logistical regression, we model predicted probabilities. In order to figure out how well our model is predicting, a ROC-curve can be helpful (Receiving Operating Characteristics). In logistical regression we want a model that predicts the outcome of the independent variable correctly at all times. The ROC-curve shows how well my model predicts by determining the relationship between true positive values (the predictions my model predicts as 1 that is observed to be 1) and false positive values (the prediction my model predicts as 1 but is actually 0) using various cut-off values. I will use the *test data* to create a ROC-curve in order to see how my models predicts on new and unseen data. Then I will compare the true positive values in the test data with the true positive values in the train data to evaluate my model's performance. I do this to see if my data is overfitted and to evaluate the overall performance of my model.

One of the ROC-curve's strengths is that it actually checks how well the model predicts at various cut-off values, and thus defines the optimal cut-off value for me. The matrix below shows the different possible outcomes.

	Vote to leave EU observed 1	**Vote to remain in EU observed 0**
Vote to leave EU predicted 1	True positive	False positive
Vote to remain in EU predicted 0	False negative	True positive

Table 2.0: This table displays the possible outcomes of my model's prediction power. Ideally the model would predict Y=1 when Y=1 is observed and Y=0 when Y=0 is observed.

The optimal point on the ROC-curve is (FP, TP) = 0.1 – no false positive predictions and all true positive predictions. Area under the curve, AUC-value, equal to 1 means that the model makes perfect predications, meaning that the model predicts Y=1 when Y=1 is observed and vice versa in all incidents.

The AUC-value for model 1 using the train (in-data-sample) data is 0.72, whilst for model 2 it is 0.77. Indicating that model 2 classifies correctly in approximately 77 percent of all incidents, whilst model 1 classifies correctly in 72 percent of all incidents.

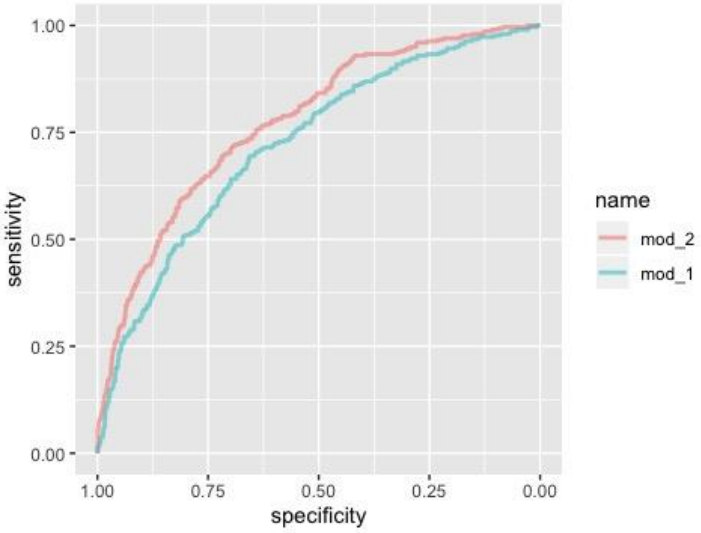
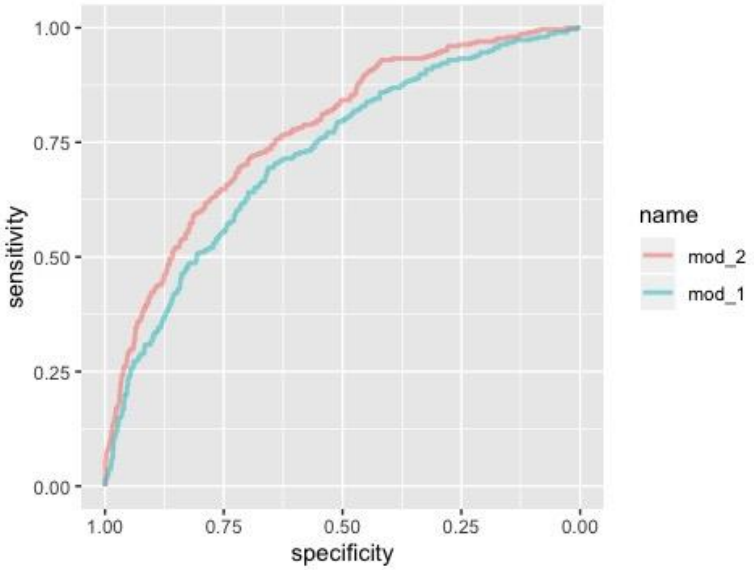


Figure 3.0: This is a figure of the ROC-curves for model 1 and 2 based on the train data. The AUC for model 1 is 0.7227, and the AUC for model 2 is 0.7741.

The ROC-curves made with the testing data has only a slight difference in predictive power, indicating that my model is good at predicting on out-of-sample-data. Model 2 is able to distinguish between Y=1 and Y=0 in approximately 77 percent of all incidents in the *test data*. Whilst model 1 predicts correct in 70 percent of all incidents in the test data. Model 2 performs just as well on new and unseen data as it does on the data that was used to make the model. Conclusively, the overall performance evaluation is that my model predicts well on new data.



Figue 6.0: The AUC-curve displays how well my model predicts in out-of-data-sample. The AUC for model 1 is 0.6997 and the AUC for model 2 is 0.7682.

Which predictions my model classifies incorrect is important to determine the overall performance of the model. Using the cut-off value from the ROC-curve for model 2 I find that

model 2 predicts Y=0 when Y = 1 in 25 percent of all incidents in the test data and approximately 22 percent of all incidents in the train data. Whilst it predicts Y = 1 when Y = 0 in 0.24 percent of all incidents in the train data and 1.4 percent of all incidents in the test data. Meaning that the model is not exactly great at predicting *would vote to leave the EU*. My model finds it difficult to predict Y = 1 when Y = 1 actually is observed, whilst it is good at classifying Y = 0. See appendix for overview of predictions.

Discussion – validity

I will in the following section briefly discuss the findings of this paper using Cook and Campbell's system of validity. According to Cook and Campbell there are four types of validity (Lund, 2002, p. 105)

The first type of validity is statistical validity. In order to meet this requirement a statistical analysis has to be able to conclude that the correlation between the independent and the dependent variable is significant (ibid.). This requirement is met as the estimated coefficients in both models are significant, except for gender in model 2. Additionally, the models fulfill the underlying assumptions of logistical regression, indicating strong statistical validity (ibid., p. 114).

The second type of validity is internal validity and refers to whether or not it is possible to conclude that the relationship between the independent and dependent variable(s) are causal (ibid., p. 105). To conclude that bad economic outlook leads to EU-skepticism is difficult when using data from one time period. The coefficients estimated shows a significant correlation, but I cannot conclude that EU-skepticism is caused by feelings of bad economic outlook. Data on the same observations over different time periods would allow for better investigation of a causal relationship between economic outlook and EU-skepticism.

The third type of validity is construct validity. Construct validity refers to the operationalization of the different variables and whether or not we measure what we seek to measure (Lund, 2002, s. 105). In this paper a *radical* measure is used to try and capture EU-skepticism. The argument for using *would vote for France to remain or leave the EU*, is to ensure strong construct validity. Arguably this variable separates the strongest EU-skeptics from the mediocre and the not-EU-skeptic, as people who are unsure and people who are not EU-skeptic would not vote leave. Using this *radical* measure of EU-skepticism offers no room for "inbetweeners." This in one way ensures that I am measuring EU-skepticism, but at the same I am not recognizing that EU-skepticism does not necessarily mean a vote to leave the EU.

Using only one indicator to measure a complex concept such as attitudes towards immigration does limit the construct validity of this paper. I have done this to allow for a parsimonious model. Nevertheless, only using one variable means that I am only measuring one aspect of anti-immigration attitudes. The same accounts for the economic explanatory variables. In an extended paper including indexes will allow for better estimation of anti-immigration attitudes as well as economic outlook.

The fourth type of validity is external validity and refers to the generalization of the results (ibid., p. 105). The results of this paper are generalizable to the French population. Firstly because of how the sample was drawn. Secondly, by looking at how well my model predicts on out-of-data sample. Nevertheless, I do not think that this model would predict well on other countries. The main reason for this is because the same estimates is not applicable to other countries because national context is important in terms of determining economic outlook and EU-skepticism. The sample of this paper is drawn to represent the population of France and not for example Germany, making it difficult to predict EU-skepticism in Germany using French observations. Nevertheless, the findings of this paper rests upon theory which does not conclude that the theory is context dependent, meaning it would be interesting to see if the same correlations is found in other countries.

Determining whether or not the results are generalizable to other times is difficult. It is possible that in the future other factors (than economic ones) may determine EU-skepticism. There are already many features determining EU-skepticism and other features might become more important in the future making economic factors (and anti-immigration attitudes) less important.

Literature

Christophersen K.A. (2013), *Introduksjon til statistisk analyse.* 1. utgave, 1. opplag, Gyldendal Akademisk, Oslo.

European Social Survey (2016), *ESS Codebook, ESS8-2016 ed.* 1.0. Available from: <http://www.europeansocialsurvey.org/docs/round8/survey/ESS8_appendix_a7_e01_0.pdf> (Downloaded: 05.09.19)

Halikiopoulou D., Nanou K. & Vasilopoulou S. (2012) The paradox of nationalism: The common denominator of radical right and radical left Euroscepticism. *European Journal of Political Research,* 2012, 51, p. 504-539.

Hermansen S.S.L. (2019), *Lær deg R. En innføring i statistikkprogrammets muligheter*. 1. utgave, 1. opplag, Fagbokforlaget, Oslo.

Hooghe L. & Marks G. (2005), *Calculation, Community and Cues,* Sage Publications, London.

Lund T. (2002), *Innføring i forskningsmetodologi.* Unipub, Oslo.

Stock J.H. & Watson M.W. (2005), *Introduction to Econometrics.* 3. Edition, Global edition, Pearson Education Limited, Essex, England.

Wuyts C. & Loosveldt G. (2019), *Quality matrix for European Social Survey, Round 8: Overall fieldwork and quality report.* Katholeike Universiteit, Leuven. Available from: < http://www.europeansocialsurvey.org/docs/round8/methods/ESS8_quality_matrix.pdf > (Downloaded: 07.10.19)

Appendix

Stats and figures

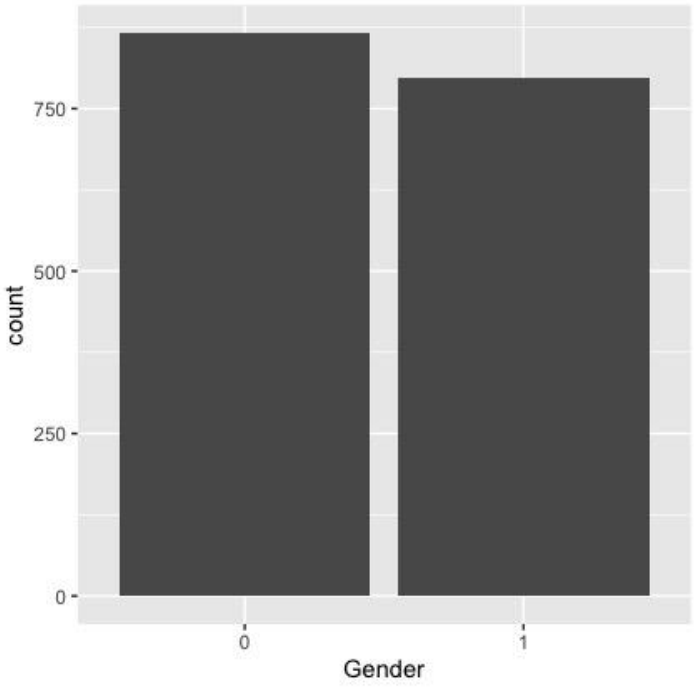
Table 1: Summary of continuous variables.

Summary

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
money	1,663	1.101	0.866	0	0	2	3
eduyrs	1,663	12.755	3.783	0	10	15	26
economy	1,663	6.514	1.939	0	5	8	10
immigration1	1,663	5.107	2.206	0	4	6	10
agea	1,663	52.282	17.910	16	38	66	99

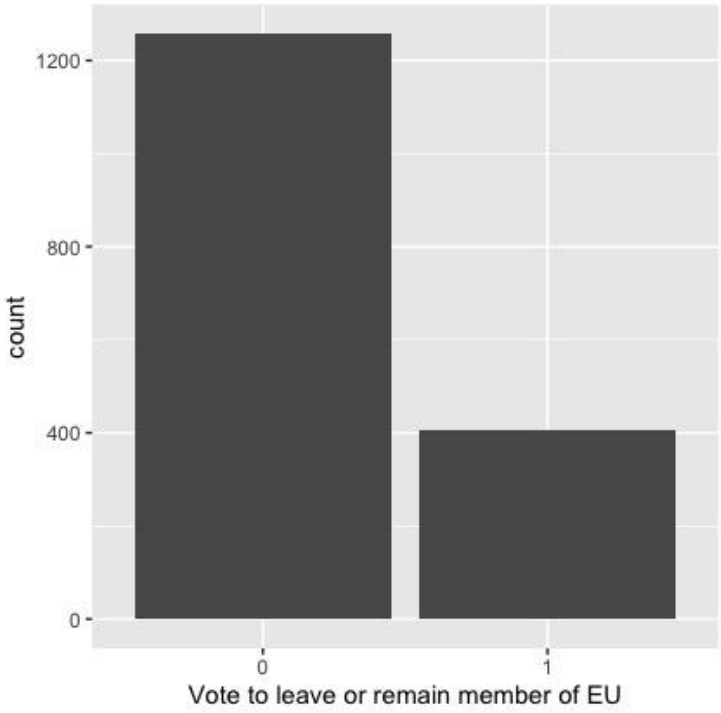
The table show the summary of each continuous variable included in the analysis. All variables have N = 1663, and it is these observations that has been separated into train data and test data.

Figure 1: Distribution of binary variable gender.



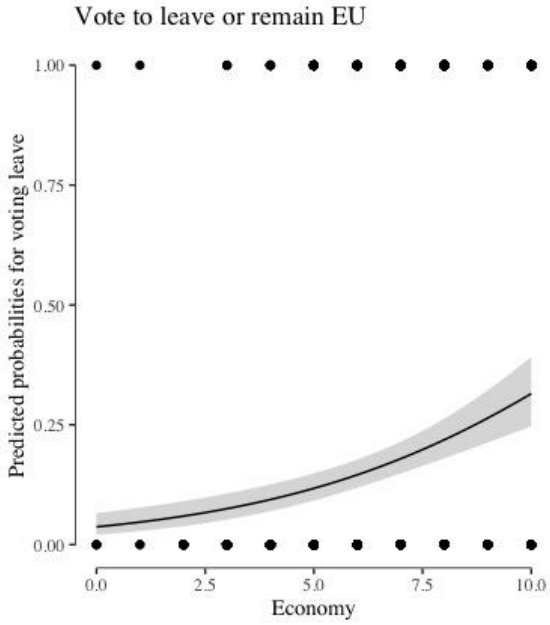
This figure shows the number of men and women that make up the observations in the data. Man is coded to 1 and women is coded to 0. There are 866 women and 797 men.

Figure 2: Distribution binary variable: *Would vote for France to leave or remain a member of the EU.*



This table shows the number of observations for vote leave = 1 and vote remain = 0. 1257 observation would vote remain and 406 would vote leave.

Figure 3: Predicted probabilities of vote to leave EU



This table shows the predicted probabilities of Y = 1, would vote for France to leave the EU. The variables that vary is economy, whilst the other variables have been coded to the median value. This table also shows that there is no complete separation.

Table 2: Correct and wrongful predictions in-data sample model 2

IN-DATA SAMPLE	**Vote to leave EU observed 1**	**Vote to remain in EU observed 0**
Vote to leave EU predicted 1	True positive = 20	False positive = 3
Vote to remain in EU predicted 0	False negative = 278	True positive = 946

This table shows the correct and wrongful predictions made by model 2 predicting on the data that was used to make the model. The cut-off value is the AUC estimated for model 2 on the in-data sample. AUC = 0.7741.

Table 3: Correct and wrongful predictions out-of-data sample model 2

OUT-OF-DATA SAMPLE	**Vote to leave EU observed, 1**	**Vote to remain in EU observed, 0**
Vote to leave EU predicted, 1	True positive = 4	False positive = 6
Vote to remain in EU predicted, 0	False negative = 104	True positive = 302

This table shows the correct and wrongful predictions made by model 2 predictions on the out-of-data sample – the data that was not used to make the model. The cut-off value for these predictions are from the AUC estimated for model 2 in the out-of-data sample. AUC= 0.7682.

R-script[4]

```
# This script is organized as follows:
# 1. European Social Survey download
# 2. Recoding variables - vote to leave EU
## 2.1 Immigration1
## 2.2 Money
## 2.3 Economy
## 2.4 Gender
## 2.5 Age
## 2.6 Years of education
# 3. Split data set
# 4. Logistical regression
## 4.1 Caret model
# 5. Calculating results from logit to probabilities
# 6. Data visualization: predicted values X on Y
# 7. Key assumptions
# 8. Goodness of fit tests:
## 8.1 Log likelihood
## 8.2 Pseudo R^2
## 8.3 Hosmer-Lemeshow-test
# 9.0. How well is my model at predicting
## 9.1 ROC Curve
## 9.2 Predicting out of data
```

[4] In order to make it easy to replicate the findings in this paper the entire R-scrips has been copied and pasted into this appendix. Be sure to set your own e-mail address when downloading the dataset.

```
library(devtools)
library(ggplot2)
library(tidyverse)
library(moments)
library(corrplot)
library(psy)
library(stargazer)
library(caret)
library(ggpubr)
library(pROC)
library(ggthemes)


##############################
###### 1.0 ESS DOWNLOAD ######
##############################

library(essurvey)

data <- import_country("France", 8,
                       ess_email = "XXXXXXXXXXXXXXXXXXX",
                       format = "stata")

recode_missings(data)


#####################################################################
###### 2.0 RECODING VOTE TO LEAVE OR REMAIN MEMBER OF EU ######
#####################################################################

# Vote EU: Remain member of EU - vteurmmb --> vteu
# 1 = Remain member of EU --> 0
# 2 = Leave the EU --> 1
# Changing values in variable to 0 and 1
data$vteu <- ifelse(data$vteurmmb == 2, "1",
                    ifelse(data$vteurmmb == 1, "0", data$vteurmmb))

table(data$vteu)
table(data$vteurmmb)


# Removing all those who do not have the values 1 or 0.
data$vteu <- ifelse(data$vteu > 11, NA, data$vteu)

# Making the variable numering
data$vteu <- as.numeric(data$vteu)
class(data$vteu)

table(data$vteu)
# 445 votes to leave
```

```
# 1370 votes to remain


####################################
###### 2.1 RECODING IMMIGRATION ######
####################################

show(data$imwbcnt)

# Changing direction of scale,
# After change : 10 = worse place to live
# 0 = better place to live

data$immigration1 <- data$imwbcnt*-1+10

table(data$imwbcnt)
table(data$immigration1)

data$immigration1 <- as.numeric(data$immigration1)

# Removing those who do not have values 0-10

data$immigration1 <- ifelse(data$immigration1 > 11, NA, data$immigration1)


#################################
###### 2.2 RECODING MONEY ######
#################################

# How likely not enough money for household necessities next 12 months
# Making sure the variable starts at 0
# Range 0-3: 0 = not likely, 3 = very likely

data$money <- ifelse(data$lknemny == "1", "0",
                     ifelse(data$lknemny == "2", "1",
                            ifelse(data$lknemny == "3", "2",
                                   ifelse(data$lknemny == "4", "3",
                                          data$lknemny))))

# Class
data$money <- as.numeric(data$money)
class(data$money) # = numeric

table(data$money)

data$money <- ifelse(data$money > 11, NA, data$money)

#################################
###### 2.3 RECODING ECONOMY ######
#################################
```

```
# How satisfied with present state of economy in country

show(data$stfeco) # values ranging from 0-10
# 0 = Extremely satisfied
# 10 = Extremely dissatisfied

data$economy <- data$stfeco*-1+10

table(data$economy)
table(data$stfeco)
# Class
class(data$economy) # = Numeric
data$economy <- as.numeric(data$economy)

data$economy <- ifelse(data$economy > 11, NA, data$economy)


###################################
###### 2.4 RECODING GENDER ######
###################################
# 1. gndr : Gender

# Overview, distribution
table(data$gndr)
show(data$gndr)

# Changing values from 1 and 2 to 0 and 1
# 1 = male = 953
# 0 = female = 1117

data$gender <- ifelse(data$gndr == 1, "1",
                      ifelse(data$gndr == 2, "0",
                             data$gndr))

# Removing values that cannot be identified as 0 or 1

data$gender <- ifelse(data$gender > 3, NA, data$gender)

show(data$gndr)
table(data$gender)
table(data$gndr)
class(data$gender) # = character
data$gender <- as.numeric(data$gender)
class(data$gender) # = numeric

###################################
###### 2.5 (RECODING) AGE ######
###################################
```

```
table(data$agea)
show(data$agea)
class(data$agea)

data$agea <- as.numeric(data$agea)
table(data$agea)

#####################################
###### 2.6 YEARS OF EDUCATION ######
#####################################

show(data$eduyrs)
table(data$eduyrs)
class(data$eduyrs)
data$eduyrs <- as.numeric(data$eduyrs)

data$eduyrs <- ifelse(data$eduyrs > 26, NA, data$eduyrs)
table(data$eduyrs)

########################################
###### 3.0 SET SEED AND SPLIT DATA ######
########################################
# New dataset, no NA.
data1 <- data %>%
  drop_na(vteu, money, eduyrs, economy, immigration1, gender, agea)

# New dataset for summary of stats
nyedata <- data1 %>%
  select(money, eduyrs, economy, immigration1, agea)

# Summary statistics table
stargazer(as.data.frame(nyedata), type = "html",
          out = "Statistic_2.html",
          title= "Summary_statistics")


# Frequency Gender

ggplot(data = data1) +
  geom_bar(mapping = aes(x = as.factor(gender))) +
  xlab("Gender")


# Frequency VOTE EU

ggplot(data = data1) +
  geom_bar(mapping = aes(x = as_factor(vteu))) +
  xlab("Vote to leave or remain member of EU") +
  title(main = )
```

```
######################### SET SEED
# Before running the regression
# I make one test dataset and one train
# I use the train dataset in the model
# The testing data set is saved to see how well my model predicts

set.seed(24)

train <- sample_frac(data1, 0.75)
testing <- data1 %>%
  anti_join(train)

#######################################
###### 4.0 LOGISTICAL REGRESSION ######
#######################################

# Model 1 - Without control variable immigration
gm4 <- glm(vteu ~ agea + gender + eduyrs + economy + money,
         data = train, family = binomial(link = "logit"))
summary(gm4)

# Model 2 - With control variable immigration

gm5 <- glm(vteu ~ agea + gender + eduyrs + economy + money + immigration1,
         data = train, family = binomial(link = "logit"))
summary(gm5)


stargazer(gm4, gm5, type = "text", title = "Results", align = T)

# More detailed table for paper
stargazer(gm4, gm5, type = "html",
        out= "Figure_1.html",
        title="Regression Results", align=TRUE,
        dep.var.labels=c("Would vote for France to remain a member of European Union or
leave"),
        covariate.labels=c("Age",
                           "Man",
                           "Years of education",
                           "Satisfaction with economy",
                           "Likely not enough money",
                           "Anti-immigration attitudes"), no.space=TRUE)



#####################################
###### 5.0 CALCULATING RESULTS ######
#####################################
# RESULTS GM5
# Calculating from logodds to oddsratio
```

```
summary(gm5)
coef(gm5) # Shows all my coefficients

#If you are a man the odds of voting leave the EU
# increases with 29 % - ceteris paribus

# Effect of economy on vote leave
(exp(coef(gm5)[5])-1)*100 #28

# If "dissatisfaction" with the economy increases with 1 unit on the scale
# the odds of voting to leave the EU increases with 28%, ceteris paribus


# Effect of money on vote leave %-change
(exp(coef(gm5)[6])-1)*100 # 42%

(exp(coef(gm4)[6])-1)*100 #results model 4
(exp(coef(gm4)[5])-1)*100 #results model 4
# If the likelihood of not having enough money in the next 12 months
# increases with 1 unit on the scale, the odds of voting to leave the EU
# increases with 42%, ceteris paribus.


# Effect of immgiration on vote leave in %-change
(exp(coef(gm5)[7])-1)*100 #38%

# If "thinking" that immigrants make the country worse increases with
# 1 unit on the scale, the odds of voting to leave the EU icreases with
# 38%, ceteris paribus.

########################################
###### 6.0 PREDICTED PROBABILITIES ######
########################################

data_for_prediction <- tibble(
  agea = median(train$agea),
  gender = median(train$gender),
  eduyrs = median(train$eduyrs),
  money = median(train$money),
  economy = seq(min(train$economy),
            max(train$economy), .1),
  immigration1 = median(train$immigration1))

predicted_data <- predict(gm5, newdata = data_for_prediction, type = "link",
                          se = T)

plot_data <- cbind(predicted_data, data_for_prediction)
```

```
plot_data$low  <- exp(plot_data$fit - 1.96*plot_data$se)/(1 + exp(plot_data$fit -
1.96*plot_data$se))
plot_data$high <- exp(plot_data$fit + 1.96*plot_data$se)/(1 + exp(plot_data$fit +
1.96*plot_data$se))
plot_data$fit <- exp(plot_data$fit)/(1+ exp(plot_data$fit))

p <- ggplot(train, aes(x = economy, y = vteu)) +
  geom_rangeframe() +
  ggtitle("Vote to leave or remain EU")  +
  theme_tufte() +
  ylab("Predicted probabilities for voting leave") +
  xlab("Economy") +
  geom_point() +
  geom_ribbon(data = plot_data, aes(y=fit, ymin=low, ymax=high), alpha=.2) +
  geom_line(data = plot_data, aes(y=fit))
p

library(ggthemes)
library(ggpubr)

###################################
###### 7.0 KEY ASSUMPTIONS ######
###################################
# 1. The regression has the shape of an -S-

preds <- predict(gm5, train, type = "response")

newdata1 <- train %>%
  select(economy, money,
         agea, immigration1, eduyrs)

newdata1 <- newdata1 %>%
  mutate(logit = log(preds/(1-preds))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)


ggplot(newdata1, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")


#######################
# 2. Influential values

# Cook's distance
plot(gm5, which = 4, id.n = 3)

# Computing standardized residuals and Cook's D
```

```
# install.packages("broom")
library(broom)
library(tidyverse)
# Model results:

model.data <- augment(gm5) %>%
  mutate(index = 1:n())

model.data %>% top_n(3, .cooksd)

ggplot(model.data, aes(index, .std.resid)) +
  geom_point(aes(color = vteu), alpha = .5) +
  theme_bw() +
  ylab("Standardized residuals") +
  xlab("Observations")

model.data %>%
  filter(abs(.std.resid) > 3)



# We have no influential observations

############################
# 3. No multicollinearity

library(car)

car::vif(gm5)

# None of my variables have a VIF-value above 2.


###################################
###### 8.0 GOODNESS OF FIT ######
###################################
### ANOVA

anova(gm4, gm5, test = "LRT")

### PSEUDO R^2
pR2(gm5)
pR2(gm4)

### HOSMER-LEMESHOW TEST
hl <- hoslem.test(gm5$y, fitted(gm5), g = 10)
hoslem.test(gm5$y, fitted(gm5), g = 10)

cbind(hl$expected, hl$observed)
```

```
summary(hl)

hl1 <- hoslem.test(gm4$y, fitted(gm4), g = 10)
hoslem.test(gm4$y, fitted(gm4), g = 10)

cbind(hl1$expected, hl1$observed)

summary(hl1)


##################################################
###### 9.0 HOW WELL DOES MY MODEL PREDICT? ######
##################################################
library(pROC)


mod5_train_roc <- roc(response = train$vteu,
                      predictor = predict(gm5, train))
mod4_train_roc <- roc(response = train$vteu,
                      predictor = predict(gm4, train))

# R-curve for train data
ggroc(list("mod_2" = mod5_train_roc,
           "mod_1" = mod4_train_roc),
      alpha = 0.5,
      linetype = 1,
      size = 1)

## AUC

auc(mod5_train_roc)
auc(mod4_train_roc)

## ROC-curve test data
mod5_test_roc <- roc(response = testing$vteu,
                     predictor = predict(gm5, testing))
mod4_test_roc <- roc(response = testing$vteu,
                     predictor = predict(gm4, testing))

ggroc(list("mod_2" = mod5_test_roc,
           "mod_1" = mod4_test_roc),
      alpha = 0.5,
      linetype = 1,
      size = 1)

auc(mod4_test_roc)
auc(mod5_test_roc)

# Confusion matrix to find which predictions my model predicts wrong
```

```
preds_test <- predict(gm5, testing, type =  "response")
preds_train <- predict (gm5, train, type = "response")



cm_test <- table(predikert = ifelse(preds_test > 0.7682, 1, 0),
                 faktisk = testing$vteu)
cm_test

cm_train <- table(prediket = ifelse(preds_train > 0.7741, 1, 0),
                  faktisk = train$vteu)

cm_train
```