FINAL DATA ANALSYSIS

An Analysis of the Voting Behaviour of the United Kingdom's General Elections in 2010

After 13 years in opposition in the United Kingdome (UK) the British Conservative party ("Tories") re-entered the government in 2010. This paper analyses the electorate of the Conservative party by answering the following question: who are the people voting for the Tories? Therefore, the paper builds on a dataset constituting of a survey of 3512 randomly selected individuals living in the UK between 18 and 90 years old.

1. Theory and Model Development

This paper analyses the electorate of the Conservative party at the 2010 general elections in the UK. The hypothesis: the more politically right individuals identify themselves on right-left self-placement scale, the more likely they will ever vote for the Conservative party in the UK. From this arises H_0: the self-placement of individuals on a right-left scale has no impact on their likelihood to ever vote for the Conservative party in the UK.

Therefore, the respondents' self-placement on a right-left scale (*rile*) from 0 (left) to 10 (right) is the main independent variable, which is expected to explain most of the variation of the dependent variable, the likelihood of ever voting for the Conservative party (*con_like*), measured on a scale form 0 (very unlikely) to 10 (very likely). Accordingly, individuals who categorize themselves predominantly right are expected to vote more likely for the respective party.

Nonetheless, since additional mechanisms may influence people's voting behaviour, the model will to control for the following variables:

- age (*age*, numerical): young people are expected more liberal and will therefore vote less likely for the Tories.
- gender (*female*, dummy variable): men tend to be more conservative and simultaneously more prone to extremist ideas. They are expected to vote more likely for conservative parties than women.
- annual household income (*income,* ordinal): individuals with a higher income will more likely to vote for the Tories because they rather represent their (economic) interests.

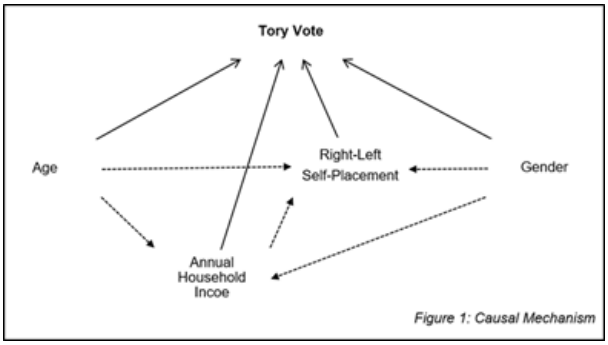


Figure 1: Causal Mechanism

Figure 1 displays the expected relationship between the five variables of the causal model. The model explicitly excludes the independent variable respondents' years of education, because most of its explanatory power will be caught by other variables.

In line with the paper's model, it will analyse the following Regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

with x_1 right-left self-placement, x_2 age, x_3 gender and x_4 annual household income. This relates to the following equation:

$$\text{con\_like} = \beta_0 + \beta_1 \text{rile} + \beta_2 \text{age} + \beta_3 \text{female} + \beta_4 \text{icome}$$

2. Analysis

Regression Model 1

	Dependent variable:
	con_like
rile	1.266***
	(0.056)
age	0.004
	(0.008)
female	0.130
	(0.230)
income	0.110***
	(0.029)
Constant	-3.553***
	(0.611)
Observations	578
R2	0.498
Adjusted R2	0.494
Residual Std. Error	2.716 (df = 573)
F Statistic	142.056*** (df = 4; 573)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression Model 1 displays the output of the model. It shows that the right-left self-placement of the individuals (*rile*) is not just highly statistically significant at a 99.9% level at explaining the dependent variable, but also substantive significant. People who identify themselves on the right-left scale from 0 (left) to right (10) one category higher are, according to their own statement, also more likely to ever vote for the Tories: in average their answers were 1.266 categories higher on a scale from 0 (very unlikely) to 10 (very likely).

While the variable *age* is neither statistically nor substantive significant, gender (*female)* – although not statistically significant – has a substantive effect. Surprisingly women said to be 0.13 categories more likely to vote for the Tories. Last but not least, the annual household income (*income*) is statistically significant but substantially even less significant than gender, with a value of 0.11.

The constant β_0 is the y-intercept, namely the value of the dependent variable, when the independent variables take on the value 0.

In line with the findings, it is true that the individuals' self-placement on a right-left scale is the most important independent variable at explaining their likelihood ever voting for the Tories when controlling for the respondents' age, gender and annual household income.

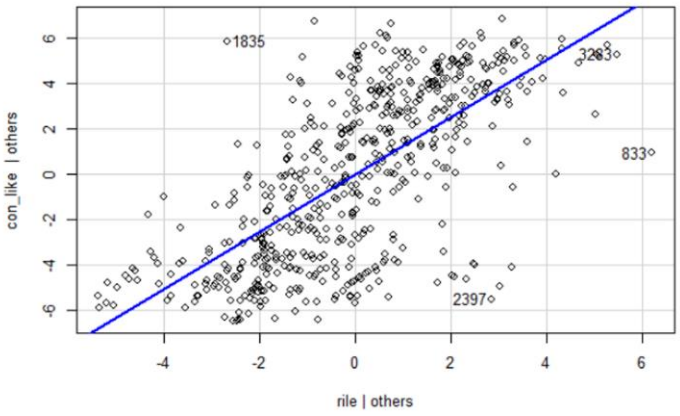
While neither age nor gender are significant, income is statistically significant and therefore, it would be interesting to explore the variable in more detail.

Regression Model 1 reports an R-squared (R^2) of 0.498 and an adjusted R^2 of 0.494. However, when dealing with Multiple Regressions one should not worry about the R^2. Only the adjusted R^2 may be helpful when comparing the explanatory power of two different models. Moreover, since the R^2 values are neither low nor high, there is no need for concern. When, instead, running a bivariate regression, containing *con_like* as the dependent and *rile* as the only independent variable, the R^2 (0.485) could be interpreted as how many percentages of the variation in the dependent variable is explained by the independent variable. However, such a underspecified model, would omitting important variables.

3. Diagnostics

Outliers are single observations with high influence (defined by their leverage * discrepancy) on a certain variable. *Graph 1* shows the outliers for the variable *rile*, namely observations labelled 833, 2397 and 1835, whereas the latter is, according to the outlier test the most influential value. However, all these observations, due to their high leverage and high discrepancy, have a higher influence on the variable then the other observations do. For a more detail analyse see additional diagnostics.

Graph 1



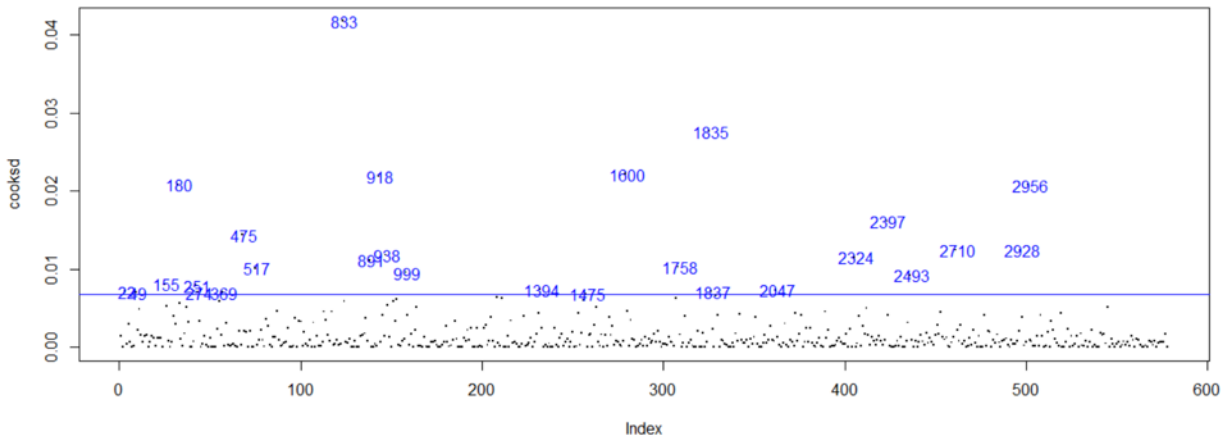
The problem with Multicollinearity is that it inflates the standard errors of the variables leading to unstable estimates, inefficiency and risking a Type II error ("false negative"), namely a non-rejection a null false null hypothesis. As shown in *Table 1*, the VIF (variation inflation factor) of the variables in the model is very low and so are the square root of the VIF ($\sqrt{\text{VIF}}$) as well as the tolerance, the invers of the VIF ($\frac{1}{\text{VIF}}$). While the former describes by how much the standard errors of the single variables are inflated, the latter shows how many

Table 1

Variable	*rile*	*age*	*female*	*income*
VIF	1.048619	1.142476	1.011194	1.136800
1/VIF	1.024021	1.068867	1.005581	1.066208
Tolerance	0.9536355	0.8752922	0.9889304	0.8796622

percentages of a certain independent variable are not explained by other independent variables. All values in *Table 1* are relatively low, Multicollinearity does not pose a risk.

When conducting an ordinary least square (OLS) regression one should test for the five OLS assumptions, namely (1) linearity, (2) mean independence, (3) homoskedasticity, (4) uncorrelated disturbances and (5) normal disturbance.
However, since the partial regression plot above already detected the outliers of one single variable and their influence on the dependent variable, it may be helpful to make use of the cook's distance to test for further outliers. *Graph 2* shows the identification number of the observations with the highest influence across all variables of the model. Additionally, the blue horizontal line represents a potential cut-off line for outliers.



Graph 2

4. EU Attitudes

As displayed in Regression Model 2 adding the respondents' attitude towards the European Union (att_eu, EU) as an independent variable influences the estimates of the other variables as well as the value of the intercept. While statistical significance did not change for neither of the variables, all of the estimates, expect for income, declined. Although not surprisingly this is due to the interactions between the newly added and the former variables.

Regression Model 2

	Dependent variable:	
	con_like	
	(1)	(2)
rile	1.266***	1.187***
	(0.056)	(0.061)
age	0.004	0.003
	(0.008)	(0.008)
female	0.130	0.069
	(0.230)	(0.230)
income	0.110***	0.124***
	(0.029)	(0.029)
factor(att_eu)2		0.502
		(0.464)
factor(att_eu)3		0.238
		(0.480)
factor(att_eu)4		-0.071
		(0.442)
factor(att_eu)5		-1.092**
		(0.520)
Constant	-3.553***	-3.108***
	(0.611)	(0.767)
Observations	578	578
R2	0.498	0.511
Adjusted R2	0.494	0.504
Residual Std. Error	2.716 (df = 573)	2.690 (df = 569)
F Statistic	142.056*** (df = 4; 573)	74.305*** (df = 8; 569)
Note:		*p<0.1; **p<0.05; ***p<0.01

.

Att_eu is a dummy variable with five categories: the reference category not displayed in the model 1 (strongly disapprove) and 2 (disapprove), 3 (neither nor), 4 (approve) and (5 strongly approve). The value of each single category must be added successively to the intercept to understand the impact of the respective category. While only the last category is statistically significant, we can observe that the most extreme values on both sides of the distribution, strongly disapprove and disapprove on the one and approve, strongly approve on the other side are substantially significant, in the opposite direction each. Accordingly, when individuals expressed to be in favour or strongly in favour of the EU, they said less likely to ever vote for the Tories, while people disapproving or strongly disapproving the EU were even more likely to ever vote for the Tories. This is expressed due to each categories impact on the constant.

	Dependent variable:	
	con_like	
	(1)	(2)
rile	1.187***	1.200***
	(0.061)	(0.061)
age	0.003	0.002
	(0.008)	(0.008)
female	0.069	0.132
	(0.230)	(0.229)
income	0.124***	0.124***
	(0.029)	(0.030)
att_eudisapprove	0.502	
	(0.464)	
att_euneither nor	0.238	
	(0.480)	
att_euapprove	-0.071	
	(0.442)	
att_eustrongly approve	-1.092**	
	(0.520)	
att_eu_int		-0.290***
		(0.107)
Constant	-3.108***	-2.189***
	(0.767)	(0.788)
Observations	578	578
R2	0.511	0.504
Adjusted R2	0.504	0.500
Residual Std. Error	2.690 (df = 569)	2.701 (df = 572)
F Statistic	74.305*** (df = 8; 569)	116.389*** (df = 5; 572)
Note:		*p<0.1; **p<0.05; ***p<0.01

Regression Model 3

The categorical variable can be transformed into a numerical data. Since, however, the variable consist of five categories, the intervals between the categories are not equal and therefore, it would be wrong to treat *att_eu* as an interval level data.

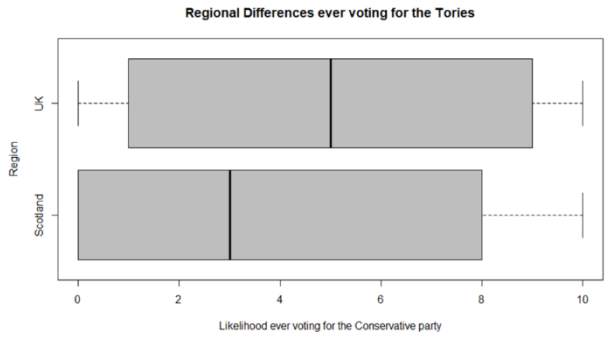
Regression Model 3 displays the differences between the categorical and numerical variable. All of a sudden, the respondents' attitude towards the EU is statistically as well as substantially significant. Simultaneously also the effect on the other independent variables varies. However, also the adjusted R^2 suggests that the explanatory power of the first linear model exceeds those of the second model. This proves that the intervals between the categories are not of equal size and that we should maintain the variable *att_eu* as a categorical level variable.

5. Region Effect

Regression Model 4

	Dependent variable:
	con_like
factor(scot)1	-0.930**
	(0.417)
Constant	5.197***
	(0.174)
Observations	578
R2	0.009
Adjusted R2	0.007
Residual Std. Error	3.807 (df = 576)
F Statistic	4.971** (df = 1; 576)
Note:	*p<0.1; **p<0.05; ***p<0.01

As displayed in *Regression Model 4*, there is a significant difference in the likelihood of ever voting for the Tories whether the respondents live in Scotland on the one or England and Wales on the other side. Accordingly, people living in Scotland responded to vote less likely for the Tories. In average their answers were approximately one category lower than people living in England or Wales. Moreover, the constant is also significant differently. Generally, although possible, dummy variables are not used as solely independent variable.



Graph 3

The boxplot *Graph 3* displays the same difference graphically. The median for people living in the UK (England and Wales) is 5, while the median for Scots is 3. Similarly, the answers of the latter are on the left side of the distribution and the normal values when it comes to the likelihood of ever voting for the Tories ranges from 0 to 8, only with a few outliers towards the right side. The values for the UK respondents instead, are more normally distributed between 1 and 9 with outliers to the left and right side of the distribution.

Conducting a difference of means test, confirms the suspicion of an actual difference in the population. Therefore, people in Scotland answered to vote less likely for the Tories than people in the UK or Wales. Below the details of the t-test conducted in R.

```
        Welch Two Sample t-test

data:  T0$con_like[T0$scot == "UK"] and T0$con_like[T0$scot == "Scotland"]
t = 2.1823, df = 142.42, p-value = 0.03072
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0875858 1.7718907
sample estimates:
mean of x mean of y
 5.197065  4.267327
```

Table 2

	mean	standard error
United Kingdom	5.197065	0.1733057
Scotland	4.267327	0.3891832

Given the t-value (> 2) and p-value (>0.05) shown in Table 2 the difference of the likelihood of ever voting for the Tories between Scotland and the UK (England and Wales) is statistically significant.

6. Interaction

Figure 2 displays the interaction effect of the variable *scot* on the relationship between the independent variable rile and the dependent variable con_like. Accordingly, column (5) of *Regression Model 5* represents the numerical values of this interaction.

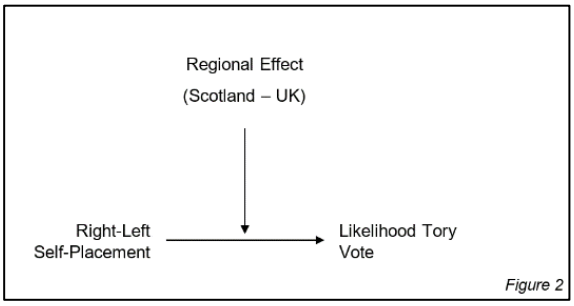


Figure 2

	Dependent variable:				
	con_like				
	(1)	(2)	(3)	(4)	(5)
rile	1.266***	1.187***	1.200***		1.282***
	(0.056)	(0.061)	(0.061)		(0.062)
age	0.004	0.003	0.002		
	(0.008)	(0.008)	(0.008)		
female	0.130	0.069	0.132		
	(0.230)	(0.230)	(0.229)		
income	0.110***	0.124***	0.124***		
	(0.029)	(0.029)	(0.030)		
att_eudisapprove		0.502			
		(0.464)			
att_euneither nor		0.238			
		(0.480)			
att_euapprove		-0.071			
		(0.442)			
att_eustrongly approve		-1.092**			
		(0.520)			
att_eu_int			-0.290***		
			(0.107)		
factor(scot)1				-0.930**	-0.773
				(0.417)	(0.846)
rile:factor(scot)1					0.023
					(0.137)
Constant	-3.553***	-3.108***	-2.189***	5.197***	-2.453***
	(0.611)	(0.767)	(0.788)	(0.174)	(0.392)
Observations	578	578	578	578	578
R2	0.498	0.511	0.504	0.009	0.489
Adjusted R2	0.494	0.504	0.500	0.007	0.486
Residual Std. Error	2.716 (df = 573)	2.690 (df = 569)	2.701 (df = 572)	3.807 (df = 576)	2.738 (df = 574)
F Statistic	142.056*** (df = 4; 573)	74.305*** (df = 8; 569)	116.389*** (df = 5; 572)	4.971** (df = 1; 576)	182.989*** (df = 3; 574)
Note:					*p<0.1; **p<0.05; ***p<0.01

Regression Model 5

While the effect is substantially significant, it lacks statistical significance. In general, residents of Scotland appear to vote less likely for the Tories. The respective regression equations for (1) UK and (2) Scottish citizens are

(1)Con_like = -2.453 + 1.282*rile

(2)Con_like = -3.226 + 1,305*rile

	Dependent variable:					
	con_like					
	(1)	(2)	(3)	(4)	(5)	(6)
rile	1.266***	1.187***	1.200***		1.282***	1.172***
	(0.056)	(0.061)	(0.061)		(0.062)	(0.067)
age	0.004	0.003	0.002			0.004
	(0.008)	(0.008)	(0.008)			(0.008)
female	0.130	0.069	0.132			0.076
	(0.230)	(0.230)	(0.229)			(0.229)
income	0.110***	0.124***	0.124***			0.122***
	(0.029)	(0.029)	(0.030)			(0.029)
att_eudisapprove		0.502				0.492
		(0.464)				(0.464)
att_euneither nor		0.238				0.205
		(0.480)				(0.480)
att_euapprove		-0.071				-0.052
		(0.442)				(0.442)
att_eustrongly approve		-1.092**				-1.103**
		(0.520)				(0.519)
att_eu_int			-0.290***			
			(0.107)			
factor(scot)1				-0.930**	-0.773	-0.930
				(0.417)	(0.846)	(0.834)
rile:factor(scot)1					0.023	0.060
					(0.137)	(0.135)
Constant	-3.553***	-3.108***	-2.189***	5.197***	-2.453***	-2.904***
	(0.611)	(0.767)	(0.788)	(0.174)	(0.392)	(0.797)
Observations	578	578	578	578	578	578
R2	0.498	0.511	0.504	0.009	0.489	0.514
Adjusted R2	0.494	0.504	0.500	0.007	0.486	0.506
Residual Std. Error	2.716 (df = 573)	2.690 (df = 569)	2.701 (df = 572)	3.807 (df = 576)	2.738 (df = 574)	2.685 (df = 567)
F Statistic	142.056*** (df = 4; 573)	74.305*** (df = 8; 569)	116.389*** (df = 5; 572)	4.971** (df = 1; 576)	182.989*** (df = 3; 574)	60.070*** (df = 10; 567)
Note:						*p<0.1; **p<0.05; ***p<0.01

Regression Model 6

The above-shown *Regression Model 6* finally represents all the regressions conducted in the course of this paper. While the first column shows the initial model, in the course of the analysis we adjusted the model and expanded it by the two variables: respondents'

attitude towards the European Union (*att_eu*) and their residence by region (*scot*). In the course the coefficient of every variable decreased, *age* remained unchanged and only *income* became more important. Simultaneously the (Adjusted) R^2 increased only marginally. Therefore, further tests concerning overspecification (i.e. Multicollinearity) should be undertaken and the model requires a clearer specification from the beginning. Now the model includes the same original variables with additional conditional determinants. This means that the strength of the relationship between the variables *rile*, *age* and *income* vary depending on the categories of the dummy *att_eu* and whether the respondents are male or female (*female*) as well as whether residents of Scotland or not (*scot*). Therefore, the model produces different regression equations.

7. Summary

	Dependent variable:
	con_like
rile	1.172***
	(0.067)
age	0.004
	(0.008)
female	0.076
	(0.229)
income	0.122***
	(0.029)
att_eudisapprove	0.492
	(0.464)
att_euneither nor	0.205
	(0.480)
att_euapprove	-0.052
	(0.442)
att_eustrongly approve	-1.103**
	(0.519)
factor(scot)1	-0.930
	(0.834)
rile:factor(scot)1	0.060
	(0.135)
Constant	-2.904***
	(0.797)
Observations	578
R2	0.514
Adjusted R2	0.506
Residual Std. Error	2.685 (df = 567)
F Statistic	60.070*** (df = 10; 567)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression Model 7

Regression Model 7 represents the final result of the regression. Although the model changed during the analysis, the respondents' self-placement on a scale from 0 (left) to 10 (right) is the most power predictor for their likelihood ever voting for the conservative party, because it is highly statistically and especially substantially significant. However, it remains quite critical by which degree the two variables *con_like* as well as *rile* differ or whether ultimately, both, have a very similar meaning. Simultaneously the value added through additional variables vary significantly and eventually the substantial significance of *income* increased, demanding further investigation.