Project in Data Anlysis: Task 5

Emilia Marchese

June 2021

1 Introduction

This project aims to analyze how the metropolitan area and the income level are related to voting behavior in Finland. The data provided on the election statistics and the income statistics in Finland will be analyzed from the viewpoint of both income variables and taxation rates. The observation variable is the municipality, represented by the variable Alue.

The task will be tackled by first providing descriptions on the data. After the data description, exploratory data analysis will be applied in order to choose the variables of interest (the political parties) and analyze the income and taxation rates variables.

The eight highest voted parties will be chosen for the data analysis and different regression models will be applied to the data. The models will be evaluated and a conclusion will be made for each of the eight parties.

2 Data description

The data sets `income2017.csv` and `ek2019.csv` were merged horizontally into a new data frame `mergedf`. The merging was done so that all the uncommon observations were dropped instead of being transformed to `NA`s in order to facilitate the data analysis. The final data set `mergedf` comprises 295 observations and 31 variables, where the observation variable is the municipality (Alue). To these 31 variables the estimated tax rates for total tax, state tax and municipal tax were added. Table 1 provides a brief description of the most important variables. For the significance of the data analysis it was decided to consider the data only for the most relevant eight political parties. The parties were picked according to the sum of the support they received in all 295 observations. This approach was chosen because most parties provide little to no information about the political behaviour of the population since they received very small and insignificant support numbers in most regions.

As can be observed from Table 1, most of the income and taxation variables are means for the respective municipality. For this reason the variable for the municipality mean income Tulot (the regional mean income) rather than Mediaanitulot (the regional median income) will be used in the data analysis, despite the fact that the mean is generally more susceptible to the presence of outliers and the median might be preferred in some cases. The estimates for the taxation rates where calculated as follows:

$$\text{VerotRate} = \frac{\text{Verot} \times 100}{\text{Tulot}}$$

$$\text{ValtionveroRate} = \frac{\text{Valtiovero} \times 100}{\text{Tulot}}$$

$$\text{KunnallisveroRate} = \frac{\text{Kunnallisvero} \times 100}{\text{Tulot}}$$

Variable	Description	Unit
Alue	Observation unit, region of Finland	Region as string
Valipiiri	Main region of Finland	Main region as string
Tulot	The average taxable income of the region's income recipients	Euros
Mediaanitulot	The taxable median income of the income recipients in the region	Euros
Ansiotulot	The average earned income of the area's income recipients	Euros
Pääomatulot	The average investement (capital) income of the area's income recipients	Euros
Verot	Total average taxes for income recipients in the area.	Euros
Valtiovero	The average state tax paid by the income recipients of the region	Euros
Kunnallisvero	The average municipal tax paid by income recipients of the area	Euros
VerotRate	Estimated tax rate for income recipients in the area from Tulot and Verot.	%
ValtioverotRate	Estimated state tax rate for income recipients in the area from Tulot and Valtionverot.	%
KunnallisveroRate	Estimated municipal tax rate for income recipients in the area from Tulot and Kunnallisverot.	%
KESK	Suomen Keskusta (Centre Party of Finland) support	%
PS	Perussuomalaiset (Finns Party) support	%
SDP	Suomen Sosiaalidemokraattinen puolue (Social Democratic Party of Finland) support	%
KOK	Kansallinen Kokoomus (National Coalition Party) support	%
VAS	Vasemmistoliitto (Left Alliance) support	%
VIHR	Vihreä liitto (Green League) support	%
RKP	Suomen ruotsalainen Kansanpuolue (Swedish People's Party of Finland) support	%
KD	Suomen Kristillisdemokraatit (Christian Democrats) support	%

Table 1: Description of relevant variables

The taxation rates were calculated because the taxation variables Verot, Valtionvero and Kunnallisvero are not as indicative of a possible relationship with party support. Therefore the data analysis on taxation will be conducted using the taxation rates variables rather than the taxation variables.

3 Data Analysis

3.1 Exploratory data analysis

Before entering deeper into the analysis of the relationship between party support and income and taxation rate variables, some relevant exploratory data analysis was conducted to get a deeper overview of the data under interest.

A summary statistics table of the parties of relevance that have been used for the data analysis is shown in Table 2. The parties are ordered in descending order according to their means. Table 3 provides summary statistics for the income and taxation in the original merged `mergedf` set.

Party	Min	1st Qu.	Median	Mean	3rd Qu.	Max
KESK	0.40	14.40	27.90	26.58	37.70	66.10
PS	1.10	16.85	19.70	19.49	22.40	42.70
SDP	2.20	10.05	15.30	15.95	21.00	40.80
KOK	0.40	6.30	10.80	11.57	16.40	32.00
VAS	0.600	3.400	5.900	7.276	9.400	28.300
VIHR	0.300	3.300	4.700	5.678	7.150	23.500
RKP	0.000	0.000	0.100	4.648	0.600	92.600
KD	0.40	2.10	3.40	4.60	5.95	48.10

Table 2: Summary statistics of the eight most relevant political parties

As can be observed from Table 2, the support of RKP stands out for having a maximum support value much higher than any other party analysed, despite having an average support rate that ranks second to last among the eight most voted parties. This possibly indicates the presence of outliers that might explain RKP anomalies. The case of the RKP support will be discussed more thoroughly in subsection 3.5.

Variable	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
Tulot	21050	23997	26164	26603	27978	63150
Ansiotulot	19090	21876	24110	24565	26173	51357
Pääomatulot	749	1614	1879	2038	2275	11793
Verot	3389	4438	5056	5286	5831	19197
Valtionvero	530	930	1124	1249	1351	10632
Kunnallisvero	2342	2900	3332	3417	3870	7460
Mediaanitulot	16403	19842	22035	22330	24106	34623

Table 3: Summary statistics of income and taxation variables

Table 3 does not provide a lot of insightful revelations on the data. One thing that could be noted is the relative proportions of income composition. It can be observed that the statistics for earned income (Ansiotulot) are much higher than the statistics for investment income (Pääomatulot), which is not quite surprising. Another point worth noting is the the difference between the third quartile values and the maximum values for most variables. As this might be indicative of a far outlier, the issue was examined more closely by plotting box plots and violin plots for the chosen variables. Box plots can easily identify outliers from the data and violin plots show the k-density distribution of the variable. These are shown in Figure 1.

What can be observed from the box plots is the presence of upper outliers for every variable. Specifically, all variables show a very far outlier with the highest income, earned income, investment income and taxes. It was analyzed whether the region that resulted in the maximum values for the variables was the same in every variable, and it was concluded that it was the same throughout. The municipality (Alue) of interest is the area of Kauniainen.

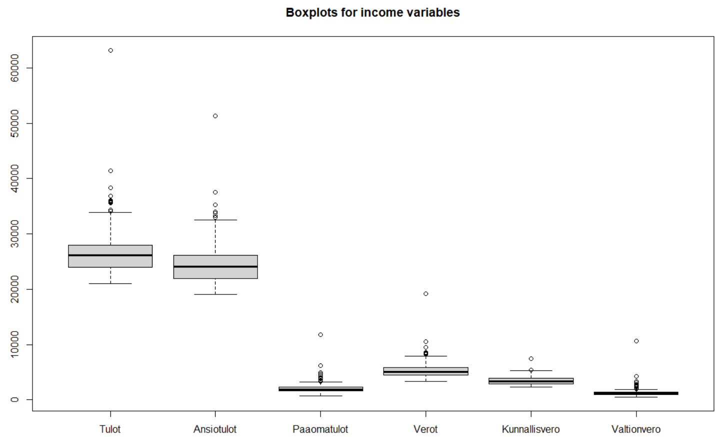
Kauniainen is a small municipality in the region of Uusiimaa, which has become attractive to high-income families due to the low municipal tax rate. The municipal tax (Kunnallsivero) was still found to be the highest for Kauniainen but its rate (KunnallisveroRate) of 11.8% was amongst the smallest municipal tax rates in the data set. The box plots and violin plots of the estimates for the tax rates are shown in Figure 2. Kauniainen remains the upper outlier in VerotRate and ValtionveroRate but as expected not in KunnallisveroRate.

It was considered whether it would be reasonable to delete this outlier from the data set and evaluate the data analysis without it. Generally, deleting outliers might cause bias in our results, but it was analyzed that in this case the values reported for the Kaunianen observation represented an exception in the distributions of most income and taxation rate variables and that keeping them might compromise the relevance of the results from the data analysis. Therefore it was deemed acceptable to not consider the observation for Kauniainen in the data analysis. All other outliers were kept in the data. With this modification, the data used for the following tables and the following sections now comprises only 294 observations.

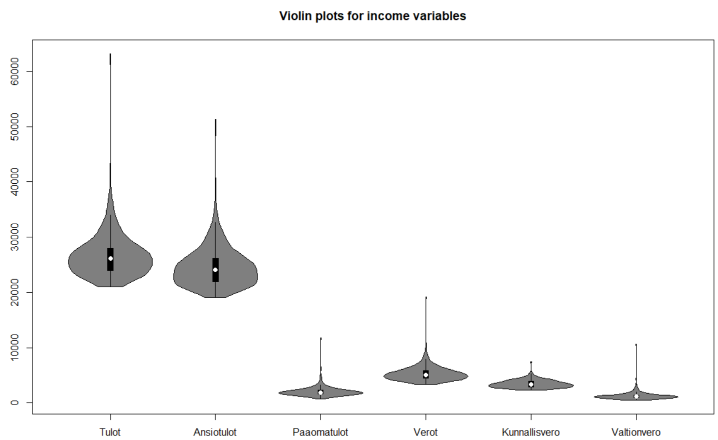
	Tulot	Ansiotulot	Pääomatulot	VerotRate	ValtionveroRate	KunnallisveroRate
Tulot	1	0.980	0.303	0.930	0.835	0.607
Ansiotulot	0.980	1	0.108	0.926	0.722	0.708
Pääomatulot	0.303	0.108	1	0.215	0.717	-0.363
VerotRate	0.930	0.926	0.215	1	0.756	0.785
ValtionveroRate	0.835	0.722	0.717	0.756	1	0.193
KunnallisveroRate	0.607	0.708	-0.363	0.785	0.193	1

Table 4: Income Correlation Matrix

Table 5 is very indicative about whether it will be expected to find a linear relationship between



(a) Boxplots



(b) Violin plots

Figure 1: Boxplot and Violin plots for variables

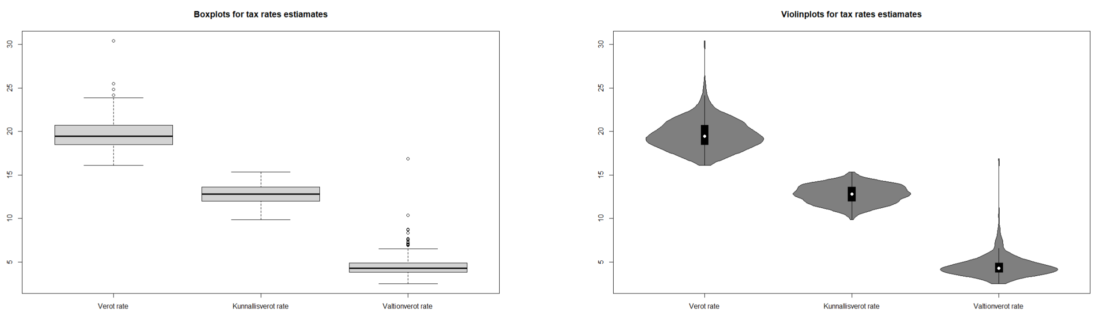
income variables and party support for the chosen party. It can be observed that the correlation values between the income variables (Tulot, Ansiotulot, Pääomatulot) and party support vary significantly between parties. Figure 3 shows a thermic correlation matrix for the values shown in Table 5.

Generally the correlation with Pääomatulot seems to be quite poor, indicating that there likely isn't a significant linear relationship between the party support and the investment income. Nevertheless this will be analyzed more thoroughly in the next section.

The correlation values (intended as absolute values) for Ansiotulot are significantly higher than the ones for Pääomatulot (investment income) for most parties, highlighting a higher chance of a possible linear relationship for the support of some parties. Accordingly, the values for Tulot are somewhere between the ones for Ansiotulot and Pääomatulot for most parties. This is not surprising as Tulot is the aggregate of Ansiotulot and Pääomatulot.

Regarding the taxation variables, we observe that VerotRate and ValtionveroRate follow a similar correlation trend to Tulot and Ansiotulot. KunnallisveroRate, on the other hand, shows slightly different correlation values than the other tax rates variables, especially with VAS and Pääomatulot.

Nevertheless, in the table there are very few values above a 0.5 correlation absolute value (by



(a) Boxplots

(b) Violin plots

Figure 2: Boxplot and Violin plots for tax rates estimates

	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD	Tulot	Ansiotulot	Pääomatulot	VerotRate	ValtionveroRate	KunnallisveroRate
KESK	1	0.252	-0.494	-0.453	0.008	-0.423	-0.463	-0.110	-0.651	-0.669	-0.044	-0.654	-0.480	-0.539
PS	0.252	1	0.044	0.150	0.039	-0.145	-0.633	-0.130	-0.077	-0.063	-0.080	-0.064	-0.125	0.029
SDP	-0.494	0.044	1	0.336	-0.123	0.252	-0.169	-0.033	0.227	0.263	-0.125	0.298	0.120	0.366
KOK	-0.453	0.150	0.336	1	-0.115	0.502	-0.273	-0.124	0.578	0.549	0.259	0.536	0.583	0.270
VAS	0.008	0.039	-0.123	-0.115	1	0.123	-0.210	-0.272	-0.016	0.048	-0.312	0.007	-0.186	0.208
VIHR	-0.423	-0.145	0.252	0.502	0.123	1	-0.163	-0.156	0.563	0.579	0.040	0.543	0.495	0.384
RKP	-0.463	-0.633	-0.169	-0.273	-0.210	-0.163	1	0.044	0.164	0.145	0.126	0.139	0.143	0.043
KD	-0.110	-0.130	-0.033	-0.124	-0.272	-0.156	0.044	1	-0.150	-0.144	-0.060	-0.134	-0.142	-0.078
Tulot	-0.651	-0.077	0.227	0.578	-0.016	0.563	0.164	-0.150	1	0.980	0.303	0.930	0.835	0.607
Ansiotulot	-0.669	-0.063	0.263	0.549	0.048	0.579	0.145	-0.144	0.980	1	0.108	0.926	0.722	0.708
Pääomatulot	-0.044	-0.080	-0.125	0.259	-0.312	0.040	0.126	-0.060	0.303	0.108	1	0.215	0.717	-0.363
VerotRate	-0.654	-0.064	0.298	0.536	0.007	0.543	0.139	-0.134	0.930	0.926	0.215	1	0.756	0.785
ValtionveroRate	-0.480	-0.125	0.120	0.583	-0.186	0.495	0.143	-0.142	0.835	0.722	0.717	0.756	1	0.193
KunnallisveroRate	-0.539	0.029	0.366	0.270	0.208	0.384	0.043	-0.078	0.607	0.708	-0.363	0.785	0.193	1

Table 5: Party Support and Income Correlation Matrix

considering only the party support correlations). These are the correlation for Ansiotulot, Tulot, Verot and KunnallisveroRate for KESK, KOK, and VIHR. Regarding ValtionveroRate only KOK shows a correlation higher than 0.5.

It might be reasonable to expect these parties to show significant values in the regression analysis that will follow. In the sections to come the support of the parties with every income and taxation rate variable will be analyzed using linear regression models in order to find possibly significant patterns for different party supports.

3.2 Simple linear regressions

It was decided to proceed by conducting regression analysis on the dependency between every party support and the income and taxation rates variables of interest. The formula that describes the regression is the simple linear regression formula:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x + \epsilon$$

Where $\hat{Y}$ is the party support, $\hat{\alpha}$ is the estimate of the constant and $\hat{\beta}$ is the parameter estimate for the coefficient, that is the slope of the best fit and the degree of change of the outcome variable every one unit change in x. For simplicity, every income variable will be expressed in hundreds rather than thousands to achieve more indicative orders of magnitudes for the parameters. This was achieved by dividing the income variables by a hundred. Therefore when x is an income variable (Tulot, Ansiotulot and Pääomatulot) the $\hat{\beta}$ will reflect a unit change in support (1%) from a 100 units change in income variable. On the other hand, when x is a taxation rate variable (VerotRate, ValtionveroRate, KunnallisveroRate), $\hat{\beta}$ reflects a unit change in support (1%) from a unit change (so 1%) in the taxation rate variable.

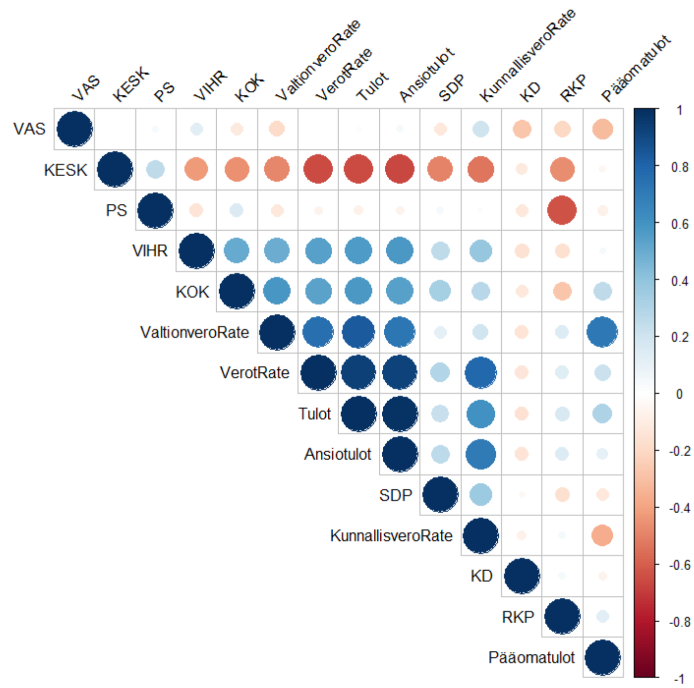


Figure 3: Correlation Matrix

Figure 4 report the scatter plots of party support and Tulot for the eight most voted parties, with the corresponding best fit line from the simple linear regression analysis. Every sub-graph shows the R^2 result from the regression analysis and the $\hat{\beta}$ parameter estimate value with the corresponding degree of significance in terms of asterisks. The significance level was reported in the graph as in the R regression summary result according to the p-value of the parameter. Table 6 provides the respective significance for every asterisks number shown in the graphs. Only the regressions with significant estimates below or equal the 5% significance level will be taken into considerations for further analysis.

Symbol	**Significance level**
$***$	0.01%
$**$	0.1%
$*$	1%
.	5%
empty	10 or greater

Table 6: Asterisks and related significance levels in scatter figures

Equal scatterplots for Ansiotulot, Pääomatulo, VerotRate, ValtionveroRate and KunnallisveroRate can be found in subsection A.1. All the regression tables are reported below. Note again that these scatter plots and the linear regressions tables do not include the outlier observation for the municipality Kauniainen. Also note that the number of asterisks used in the scatter plots correspond to the significance level expressed by Table 6, whereas the asterisks reported in the regression tables correspond to the values in the note section of each regression table.

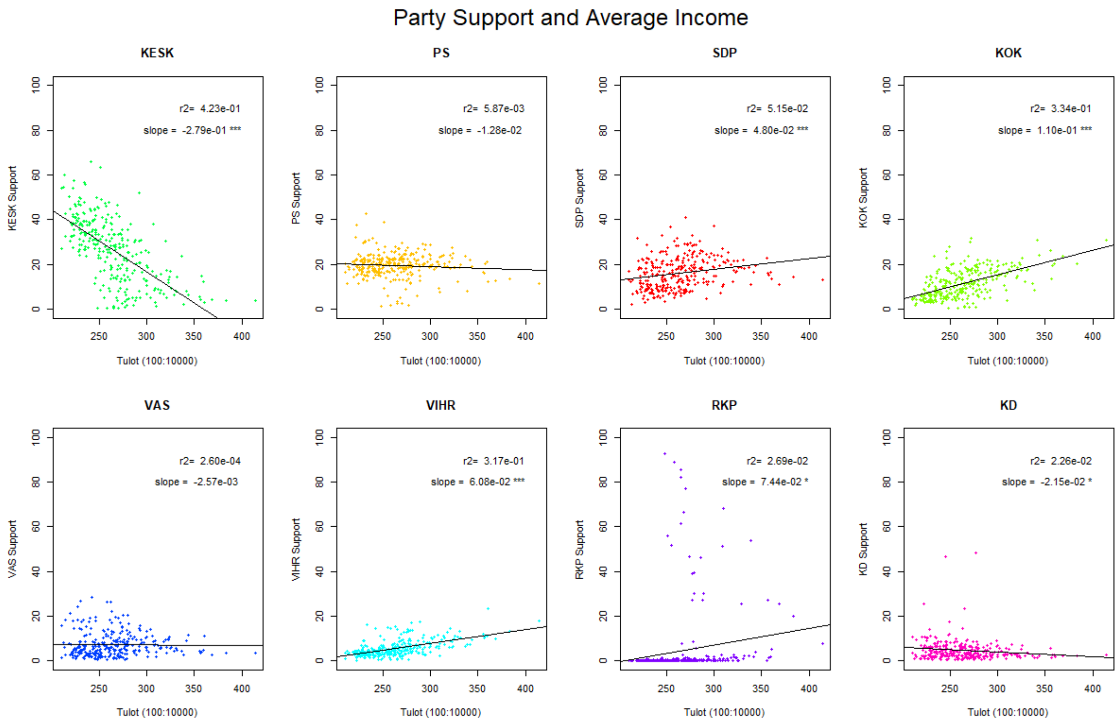


Figure 4: Party support and average total income scatter plots

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tulot/100	-0.279^{***}	-0.013	0.048^{***}	0.110^{***}	-0.003	0.061^{***}	0.074^{***}	-0.022^{***}
	(0.019)	(0.010)	(0.012)	(0.009)	(0.009)	(0.005)	(0.026)	(0.008)
Constant	100.562^{***}	22.935^{***}	3.286	-17.673^{***}	7.976^{***}	-10.435^{***}	-15.162^{**}	10.311^{***}
	(5.089)	(2.612)	(3.216)	(2.430)	(2.490)	(1.393)	(6.986)	(2.210)
Observations	294	294	294	294	294	294	294	294
R^2	0.423	0.006	0.052	0.334	0.0003	0.317	0.027	0.023
Adjusted R^2	0.421	0.002	0.048	0.332	-0.003	0.315	0.024	0.019
Residual Std. Error (df = 292)	10.965	5.627	6.928	5.236	5.366	3.001	15.052	4.762
F Statistic (df = 1; 292)	214.206^{***}	1.724	15.862^{***}	146.378^{***}	0.076	135.648^{***}	8.085^{***}	6.763^{***}
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 7: Simple linear regression with Tulot for party supports

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ansiotulot/100	-0.300^{***}	-0.011	0.058^{***}	0.109^{***}	0.008	0.065^{***}	0.069^{**}	-0.022^{**}
	(0.019)	(0.010)	(0.012)	(0.010)	(0.010)	(0.005)	(0.027)	(0.009)
Constant	99.982^{***}	22.244^{***}	1.811	-15.215^{***}	5.339^{**}	-10.295^{***}	-12.252^{*}	9.887^{***}
	(4.802)	(2.522)	(3.073)	(2.401)	(2.400)	(1.326)	(6.760)	(2.134)
Observations	294	294	294	294	294	294	294	294
R^2	0.448	0.004	0.069	0.301	0.002	0.335	0.021	0.021
Adjusted R^2	0.446	0.001	0.066	0.299	-0.001	0.333	0.018	0.018
Residual Std. Error (df = 292)	10.725	5.632	6.863	5.363	5.360	2.961	15.097	4.766
F Statistic (df = 1; 292)	237.127^{***}	1.176	21.659^{***}	125.884^{***}	0.676	147.339^{***}	6.281^{**}	6.222^{**}
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 8: Simple linear regression with Ansiotulot for party supports

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pääomatulot/100	−0.094	−0.067	-0.133^{**}	0.248^{***}	-0.250^{***}	0.022	0.286^{**}	−0.043
	(0.126)	(0.049)	(0.062)	(0.054)	(0.045)	(0.032)	(0.132)	(0.042)
Constant	28.550^{***}	20.880^{***}	18.656^{***}	6.518^{***}	12.310^{***}	5.225^{***}	−1.185	5.476^{***}
	(2.661)	(1.038)	(1.302)	(1.143)	(0.941)	(0.670)	(2.793)	(0.887)
Observations	294	294	294	294	294	294	294	294
R^2	0.002	0.006	0.016	0.067	0.098	0.002	0.016	0.004
Adjusted R^2	-0.002	0.003	0.012	0.064	0.094	-0.002	0.012	0.0002
Residual Std. Error (df = 292)	14.423	5.626	7.057	6.196	5.098	3.629	15.138	4.808
F Statistic (df = 1; 292)	0.558	1.873	4.652^{**}	21.070^{***}	31.578^{***}	0.472	4.678^{**}	1.062
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 9: Simple linear regression with Pääomatulot for party supports

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VerotRate	-5.722^{***}	−0.219	1.282^{***}	2.084^{***}	0.023	1.195^{***}	1.284^{**}	-0.390^{**}
	(0.387)	(0.200)	(0.241)	(0.192)	(0.190)	(0.108)	(0.536)	(0.169)
Constant	138.767^{***}	23.826^{***}	-9.135^{*}	-29.337^{***}	6.849^{*}	-17.758^{***}	-20.618^{*}	12.255^{***}
	(7.614)	(3.927)	(4.735)	(3.775)	(3.741)	(2.126)	(10.535)	(3.328)
Observations	294	294	294	294	294	294	294	294
R^2	0.428	0.004	0.089	0.287	0.00005	0.295	0.019	0.018
Adjusted R^2	0.426	0.001	0.085	0.285	-0.003	0.293	0.016	0.015
Residual Std. Error (df = 292)	10.921	5.632	6.791	5.415	5.366	3.050	15.111	4.773
F Statistic (df = 1; 292)	218.304^{***}	1.204	28.359^{***}	117.796^{***}	0.014	122.174^{***}	5.745^{**}	5.317^{**}
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 10: Simple linear regression with VerotRate for party supports

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ValtionveroRate	-6.326^{***}	-0.645^{**}	0.776^{**}	3.409^{***}	-0.908^{***}	1.638^{***}	1.996^{**}	-0.623^{**}
	(0.676)	(0.299)	(0.377)	(0.278)	(0.281)	(0.168)	(0.806)	(0.254)
Constant	55.010^{***}	22.424^{***}	12.513^{***}	-3.779^{***}	11.366^{***}	-1.679^{**}	−4.398	7.400^{***}
	(3.117)	(1.378)	(1.739)	(1.284)	(1.298)	(0.777)	(3.717)	(1.174)
Observations	294	294	294	294	294	294	294	294
R^2	0.231	0.016	0.014	0.339	0.034	0.245	0.021	0.020
Adjusted R^2	0.228	0.012	0.011	0.337	0.031	0.242	0.017	0.017
Residual Std. Error (df = 292)	12.662	5.599	7.062	5.214	5.273	3.157	15.101	4.768
F Statistic (df = 1; 292)	87.616^{***}	4.665^{**}	4.240^{**}	150.039^{***}	10.420^{***}	94.529^{***}	6.133^{**}	5.992^{**}
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 11: Simple linear regression with ValtionveroRate for party supports

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KunnallisveroRate	-7.374^{***}	0.157	2.464^{***}	1.642^{***}	1.055^{***}	1.321^{***}	0.627	−0.356
	(0.674)	(0.313)	(0.367)	(0.342)	(0.291)	(0.186)	(0.845)	(0.266)
Constant	120.847^{***}	17.525^{***}	-15.483^{***}	-9.478^{**}	-6.183^{*}	-11.204^{***}	−3.464	9.155^{***}
	(8.633)	(4.006)	(4.701)	(4.386)	(3.728)	(2.382)	(10.826)	(3.410)
Observations	294	294	294	294	294	294	294	294
R^2	0.291	0.001	0.134	0.073	0.043	0.147	0.002	0.006
Adjusted R^2	0.289	-0.003	0.131	0.070	0.040	0.145	-0.002	0.003
Residual Std. Error (df = 292)	12.156	5.641	6.620	6.176	5.249	3.354	15.244	4.802
F Statistic (df = 1; 292)	119.836^{***}	0.253	45.127^{***}	23.024^{***}	13.163^{***}	50.492^{***}	0.551	1.790
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 12: Simple linear regression with KunnallisveroRate for party supports

3.3 Multiple linear regressions

Multiple linear regression will give deeper insights on the relationship between variables by also accounting for the effect that other variables might have on our dependent variables. By including more than one variable in our regression we may minimize the chance of omitted variable bias (in case variables correlated to both our dependent and independent variable are excluded) and portray a more realistic model. This is unfortunately not always the case, and sometimes a simple linear model is to be preferred to the multiple linear model.

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \epsilon$$

The equation for a multiple linear regression model with two variables is reported above. Where $\hat{Y}$ is the chosen party support. The parameter estimates in the multivariate model take another interpretation from the simple linear model. For example, the $\hat{\beta}$ parameter for the X_1 variable in the simple model represents the change in the party support every unit change in the X_1 value. This parameter estimate is calculated by taking the effect of X_1 on the party support in isolation from every other variable. On the other hand, in the multiple model the parameter estimate $\hat{\beta}_1$ reflects the associated change in the party support every unit change in the X_1 variable by also taking into account all other explanatory variables in the model, in this case X_2. The $\hat{\beta}_1$ essentially reflects the change in Y from every unit change in the X_1 when the values for X_2 are kept constant. A reason to use multiple linear regression is to analyze whether a significant relationship in the simple linear model disappears in the multiple linear model. This is often the case where we use two correlated independent variables in the multiple linear model, only to discover that one of them turns insignificant from the linear model to the multiple model. This often entails that the significance shown by variable in the simple model was in reality caused by another variable that the independent variable was correlated to, rather than the independent variable itself, ultimately causing ommitted variable bias.

To show what is meant by this, it might be helpful to have a look at Tulot and VerotRate. Tulot and VerotRate are highly positively correlated to each other (see correlation matrix in Figure 3) and both simple regressions with Tulot and VerotRate (Table 7 and Table 10) return significant parameters for KESK, SDP, KOK, VIHR, RKP and KD. To analyze whether there is an actual relationship between these variables and party support, the multiple linear regression model is analyzed. The equation is given below.

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \frac{\text{Tulot}}{100} + \hat{\beta}_2 \text{VerotRate} + \epsilon$$

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tulot/100	-0.134^{***}	-0.021	-0.078^{**}	0.112^{***}	-0.027	0.047^{***}	0.118	-0.028
	(0.051)	(0.027)	(0.032)	(0.025)	(0.025)	(0.014)	(0.072)	(0.023)
VerotRate	-3.179^{***}	0.184	2.771^{***}	-0.044	0.533	0.313	-0.951	0.136
	(1.046)	(0.545)	(0.652)	(0.508)	(0.519)	(0.290)	(1.458)	(0.462)
Constant	124.437^{***}	21.556^{***}	-17.526^{***}	-17.345^{***}	3.973	-12.785^{***}	-8.019	9.288^{**}
	(9.325)	(4.860)	(5.806)	(4.523)	(4.626)	(2.587)	(12.992)	(4.113)
Observations	294	294	294	294	294	294	294	294
R^2	0.441	0.006	0.107	0.334	0.004	0.320	0.028	0.023
Adjusted R^2	0.437	-0.001	0.101	0.329	-0.003	0.315	0.022	0.016
Residual Std. Error (df = 291)	10.814	5.636	6.733	5.245	5.365	3.000	15.067	4.769
F Statistic (df = 2; 291)	114.737^{***}	0.916	17.439^{***}	72.944^{***}	0.565	68.442^{***}	4.247^{**}	3.415^{**}
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 13: Multivariate regression for party supports with Tulot and VerotRate

We observe that in the multiple regression results shown in Table 13, $\hat{\beta}_1$ and $\hat{\beta}_2$ remain both significant only for KESK and SDP. On the other hand for KOK and VIHR only $\hat{\beta}_1$ for Tulot remains significant, indicating that despite the fact that a significant relationship between VerotRate and KOK and VIHR was found in the simple models, this disappears in the multiple model. It could be reasonable to assume that in the simple model, significant results where achieved only due to the correlation between VerotRate and Tulot, leading to omitted variable bias in the case of the VerotRate simple model for KOK and VIHR.

On the other hand, the presence of both significant parameters in the multiple model for KESK and SDP indicate that both Tulot and VerotRate have an effect on the party support, which is not the result of the two variables being correlated to each other.

Another interesting multiple model to analyze would be the one with Pääomatulot and KunnallisveroRate. It can be observed from the correlation matrix in Figure 3 that these two variables show a significant negative correlation. In the simple model in Table 9, the Pääomatulot parameter is significant for SDP, KOK, VAS and RKP. The simple model with KunnallisveroRate in Table 12 shows a significant parameter for KESK, SDP, KOK, VAS, VIHR. The multiple regression model equation is shown below and the regression results can be found in Table 14.

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \frac{\text{Pääomatulot}}{100} + \hat{\beta}_2 \text{KunnallisveroRate} + \epsilon$$

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pääomatulot/100)	-0.595^{***}	-0.067	0.009	0.395^{***}	-0.219^{***}	0.112^{***}	0.371^{***}	-0.073
	(0.109)	(0.053)	(0.062)	(0.053)	(0.048)	(0.031)	(0.141)	(0.045)
KunnallisveroRate	-8.746^{***}	0.003	2.486^{***}	2.552^{***}	0.551^{*}	1.579^{***}	1.482^{*}	-0.525^{*}
	(0.690)	(0.335)	(0.394)	(0.337)	(0.302)	(0.195)	(0.898)	(0.285)
Constant	150.289^{***}	20.845^{***}	-15.950^{***}	-29.002^{***}	4.635	-16.758^{***}	-21.816^{*}	12.785^{***}
	(9.835)	(4.780)	(5.625)	(4.813)	(4.307)	(2.787)	(12.803)	(4.062)
Observations	294	294	294	294	294	294	294	294
R^2	0.357	0.006	0.134	0.220	0.108	0.185	0.025	0.015
Adjusted R^2	0.353	-0.0005	0.128	0.215	0.102	0.179	0.018	0.008
Residual Std. Error (df = 291)	11.595	5.635	6.631	5.674	5.077	3.285	15.093	4.788
F Statistic (df = 2; 291)	80.840^{***}	0.933	22.500^{***}	41.149^{***}	17.581^{***}	32.945^{***}	3.715^{**}	2.235
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 14: Multivariate regression for party supports with Pääomatulot and KunnallisveroRate

It can be observed from Table 14 that in the multiple model only KESK, KOK and VIHR have two significant parameters (ignoring the costant), indicating that for these parties Pääomatulot and KunnallisveroRate have an effect independent of each other. Actually, VIHR in the multiple model shows a significant parameter for Pääomatulot which was not significant in the simple model. This might also be related to the fact that the simple model with Pääomatulot for VIHR suffered from omitted variable bias (likely from the lack of KunnallisveroRate itself) that disappears in the multiple model once the bias is rectified by the inclusion of the correlated variable (KunnallisveroRate). Of course these are only conjectures and further analysis by integrating other variables should be made. In the multiple model, SDP shows a significant parameter only for Kunnallisvero, while VAS and RKP show a significant parameter only for Pääomatulot.

Lastly, the multiple with Ansiotulot and Pääomatulot as independent variables was analyzed. The model is described by the equation:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \frac{\text{Ansiotulot}}{100} + \hat{\beta}_2 \frac{\text{Pääomatulot}}{100} + \epsilon$$

In the simple model with Ansiotulot the Ansiotulot parameter was significant for KESK, SDP, KOK, VIHR, RKP, KD. In the simple model Pääomatulot showed a significant parameter for KOK, VAS, SDP and RKP. The multiple regression results are shown in Table 15.

	Dependent variable:							
	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ansiotulot/100	-0.301^{***}	-0.010	0.062^{***}	0.105^{***}	0.014	0.065^{***}	0.063^{**}	-0.021^{**}
	(0.020)	(0.010)	(0.012)	(0.010)	(0.009)	(0.005)	(0.027)	(0.009)
Pääomatulot/100	0.062	-0.062	-0.165^{***}	0.194^{***}	-0.257^{***}	-0.012	0.253^{*}	-0.032
	(0.094)	(0.049)	(0.060)	(0.046)	(0.045)	(0.026)	(0.132)	(0.042)
Constant	99.075^{***}	23.149^{***}	4.212	-18.038^{***}	9.085^{***}	-10.117^{***}	-15.938^{**}	10.359^{***}
	(4.999)	(2.620)	(3.160)	(2.428)	(2.369)	(1.381)	(6.998)	(2.221)
Observations	294	294	294	294	294	294	294	294
R^2	0.449	0.009	0.093	0.342	0.104	0.336	0.033	0.023
Adjusted R^2	0.445	0.003	0.087	0.337	0.098	0.331	0.027	0.016
Residual Std. Error (df = 291)	10.735	5.627	6.786	5.214	5.087	2.965	15.028	4.769
F Statistic (df = 2; 291)	118.553^{***}	1.381	14.905^{***}	75.554^{***}	16.953^{***}	73.583^{***}	5.009^{***}	3.407^{**}
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 15: Multiple regression for party supports with Ansiotulot and Pääomatulot

In the multiple model, only KOK and SDP keep both parameters significant. For these two parties it may be claimed that Ansiotulot and Pääomatulot have an individual effect on the party support. KESK, VIHR and RKP show a significant parameter only for Ansiotulot, and VAS shows a significant parameter only for Pääomatulot. The reasons these might accur are anologous to the explanations given for the previous multiple models.

Multiple regression analysis was conducted using different variables and a different number of independent variables. Only the ones that resulted in potentially significant results were picked. For the regressions with three independent variables, the number of significant parameters were not many and therefore the models were deemed not appropriate to describe any the party support.

3.4 The interaction model

In the previous subsections the party support for the simple linear model and the multiple linear model with the chosen income and taxation rate variables were analyzed.

In this section it will be discussed whether an interaction model would fit better for certain parties, more specifically KESK, VIHR and SDP. These parties were chosen because the parameter sign for an independent variable switches signs from the simple model to the multiple model (see Table 18 in section 4 for comparison).

The ambiguous switching of sign from the simple to multiple model might be indicative of the Simpson Paradox. This paradox describes the phenomenon in which trends that exist in in aggregate disappear or reverse when the data is broken into smaller groups or brackets. In order to evaluate whether our models might be experiencing such phenomena, the variables for which the parameters are switching signs should be broken into smaller groups.

It was shown that while Ansiotulot and Pääomatulot both show a positive $\hat{\beta}$ parameter in the simple linear regression for KESK (at 5% and 10% significance level respectively), the $\hat{\beta}_2$ parameter for the Pääomatulot changes to positive in the multiple model with independent variable Ansiotulot and Pääomatulot shown in Table 15. Similar observations could be made for the VIHR simple and multivariate model with Ansiotulot and Pääomatulot as independent variables.

Based on such observations and despite the fact that Pääomatulot has an insignificant parameter for both parties in the multiple model, it could be sensible to divide the Ansiotulot variable into brackets corresponding to the different Ansiotulot income levels, and then evaluate the relationship between Pääomatulot and party support for those observations in the same Ansiotulot bracket. This would correspond to the interaction linear regression equation reported below:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \frac{\text{Pääomatulot}}{100} + \hat{\beta}_2 \text{AnsiotulotBraket} + \hat{\beta}_{int} \frac{\text{Pääomatulot}}{100} \times \text{AnsiotulotBracket} + \epsilon$$

Where AnsiotulotBraket is a categorical variable resulting from the division of Ansiotulot into 4 groups ("High", "Medium-High", "Medium-Low", "Low") according to the 25%, 50% and 75% percentile. Table 16 shows the regression table of the interaction model with Pääomatulot and AnsiotulotBraket. The value for I(Pääomatulot/100) reflects the parameter estimates for those municipalities falling in the AnsiotulotBracket with value "High". Figure 5 shows a scatter plot for KESK and VIHR with the best fit lines for every AnsiotulotBraket.

	Dependent variable:	
	KESK	VIHR
	(1)	(2)
I(Pääomatulot/100)	−0.076 (0.175)	0.131** (0.051)
AnsiotulotBraketLow	28.299*** (5.885)	−3.142* (1.725)
AnsiotulotBraketMedium-High	5.472 (5.764)	2.369 (1.690)
AnsiotulotBraketMedium-Low	18.655*** (5.314)	−0.102 (1.558)
I(Pääomatulot/100):AnsiotulotBraketLow	−0.142 (0.278)	−0.060 (0.081)
I(Pääomatulot/100):AnsiotulotBraketMedium-High	0.178 (0.280)	−0.226*** (0.082)
I(Pääomatulot/100):AnsiotulotBraketMedium-Low	0.005 (0.240)	−0.186*** (0.070)
Constant	14.934*** (3.857)	5.573*** (1.131)
Observations	294	294
R^2	0.460	0.266
Adjusted R^2	0.446	0.248
Residual Std. Error (df = 286)	10.723	3.144
F Statistic (df = 7; 286)	34.755***	14.827***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 16: Interaction among Pääomatulot and Ansiotulot Braket for KESK and VIHR

Table 16 reveals that none of the parameter estimates for KESK are significant (ignoring the costants). This signifies that statistically the interaction model is not quite valid. By observing the KESK scatter plot by Ansiotulot bracket in Figure 5, it can be observed that there does not seem to be a relevant difference in the slope between the different Ansiotulot brackets. If anything, we observe what could already been observed in the scatterplot for KESK shown in Figure 9 (subsection A.1), where municipalities shows progressively low support the higher the Ansiotulot.

On the other hand, the interaction model for VIHR achieved significant parameter values for the "High" bracket, the "Medium-High" bracket and the "Medium-Low" bracket. There seems to be a positive relationship between Pääomatulot and VIHR party support for those municipalities in the "High" Ansiotulot bracket, and a negative one for those municipalities in the "Medium-High" and "Medium-Low" Ansiotulot bracket. The reasons for such relationships are quite open to speculation, if they are actually relevant. Probably some background knowledge on the VIHR party and their political aims would come in use right now.

For the SDP case, it is worth noting that in the simple model Tulot shows a positive significant parameter and VerotRate also shows a positive significant parameter. Nevertheless, in the multiple model with Tulot and VerotRate as independent variable (Table 13), the parameter estimates for Tulot switches to negative and becomes significant. Therefore, to analyze whether the interaction model could result to more relevant results, the VerotRate variables was divided into four brackets. The interaction model is expressed by the equation below:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \frac{\text{Tulot}}{100} + \hat{\beta}_2 \text{VerotBraket} + \hat{\beta}_{int} \frac{\text{Tulot}}{100} \times \text{VerotBracket} + \epsilon$$

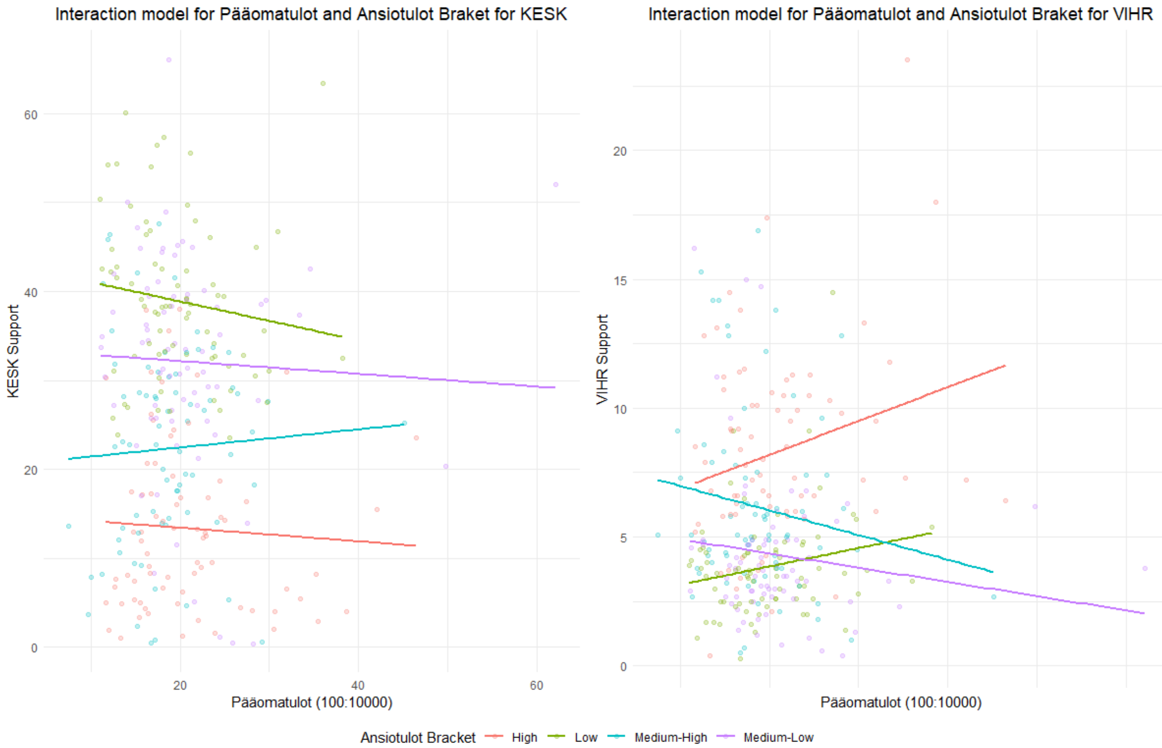


Figure 5: Interaction model for KESK and VIHR

The regression results are shown in Table 17 and a visual representation is given in Figure 6. The interaction model for SDP resulted in one significant parameter, for the "High" VeroBraket. The relationship turns out to be negative, which can be interpreted as municipalities in the "High" VerotBracket bracket showing less support for SDP as their total average income increases (Tulot). Again, some background knowledge could be useful in trying to find a reason for why this relationship is observed. These two models where picked due to the shift of sign that they showed between the simple and the multiple model, but other models that have not been conducted could have

Overall there are so many other interactions that could be explored and that could possibly reveal significant patterns in the data. The interactions model analyzed where chosen due to the sign changed anomaly that they showed between the simple and multiple models. Nevertheless, there might be other interaction models that have not been conducted that could also have revealed some insightful information on the support for some parties.

	Dependent variable:
	SDP
I(Tulot/100)	−0.064** (0.026)
VerotBraketLow	−13.076 (17.163)
VerotBraketMedium-High	−24.450 (20.377)
VerotBraketMedium-Low	−22.541 (16.611)
I(Tulot/100):VerotBraketLow	0.009 (0.070)
I(Tulot/100):VerotBraketMedium-High	0.079 (0.074)
I(Tulot/100):VerotBraketMedium-Low	0.054 (0.063)
Constant	38.629*** (8.133)
Observations	294
R^2	0.167
Adjusted R^2	0.147
Residual Std. Error	6.559 (df = 286)
F Statistic	8.204*** (df = 7; 286)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 17: Interaction among Tulot and VerotBraket for SDP

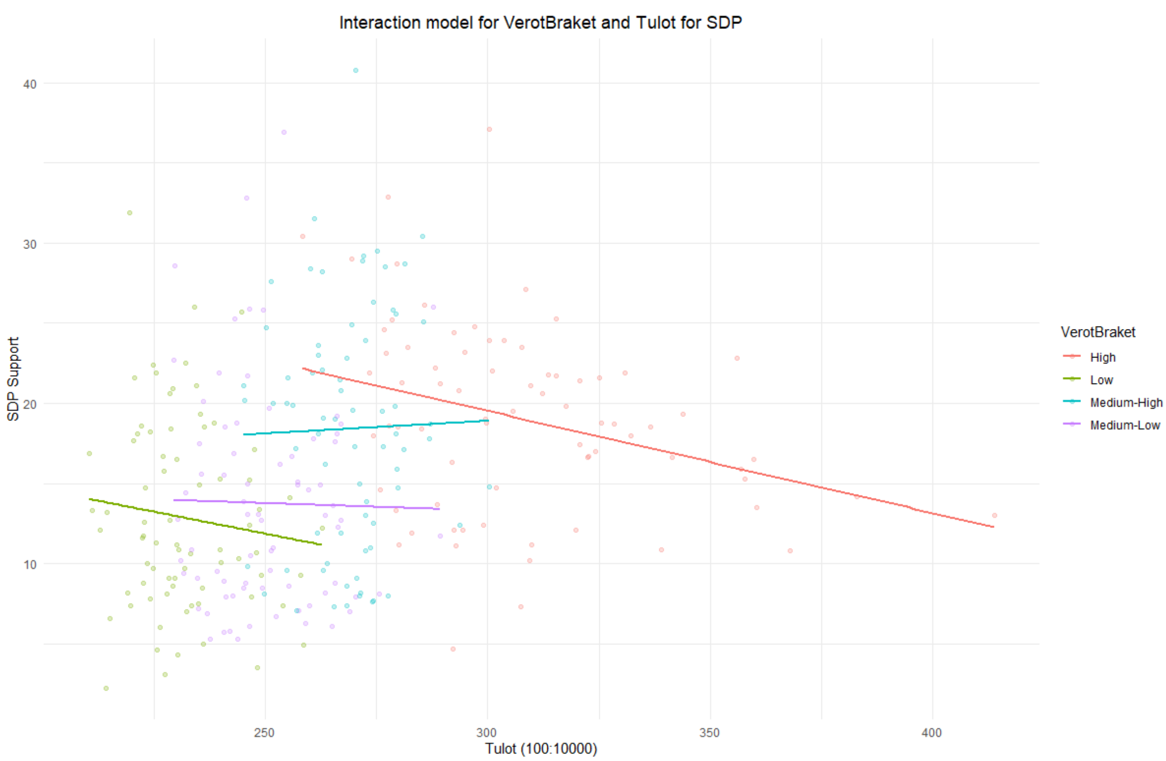


Figure 6: Interaction model for SDP

3.5 An overlook on Vaalipiiri and the RKP case

The support for the RKP party seems to not follow a particular logic. In the regression tables shown in subsection 3.2 we observe that the Tulot, Ansiotulot and Pääomatulot variables show a significant and positive relatioship with RKP support, but the regressions r squared parameter are very poor, indicating that the line of the best fit does not describe the RKP support trend quite well. This can be clearly seen by observing the scatterplots in subsection A.1. For every RKP scatter plot we observe these numerous “outliers” with extremely high RKP support numbers. In one area specifically the RKP support reaches 92.6%.

It was tried to explain the RKP trend by using multiple regression combinations of income and taxation variables as independent variables, but the R^2 value was overall poor and the models where therefore not deeply analyzed. The multiple linear regression examined in subsection 3.3 showed significant parameters but, as for the simple regression, a very poor R^2.

It should be recalled that RKP stands for Swedish Coalition Party. This said, it becomes immediately sensible to guess that those areas that show extremely high RKP support must have a significant Finnish-Swedish inhabitants. It was then decided to briefly examine the eight parties support by the main region Vaalipiiri. The variable Vaalipiiri represents the 12 regions of Finland. The scatter plots are shown in Figure 7, where the observations belonging to different regions are highlighted by color.

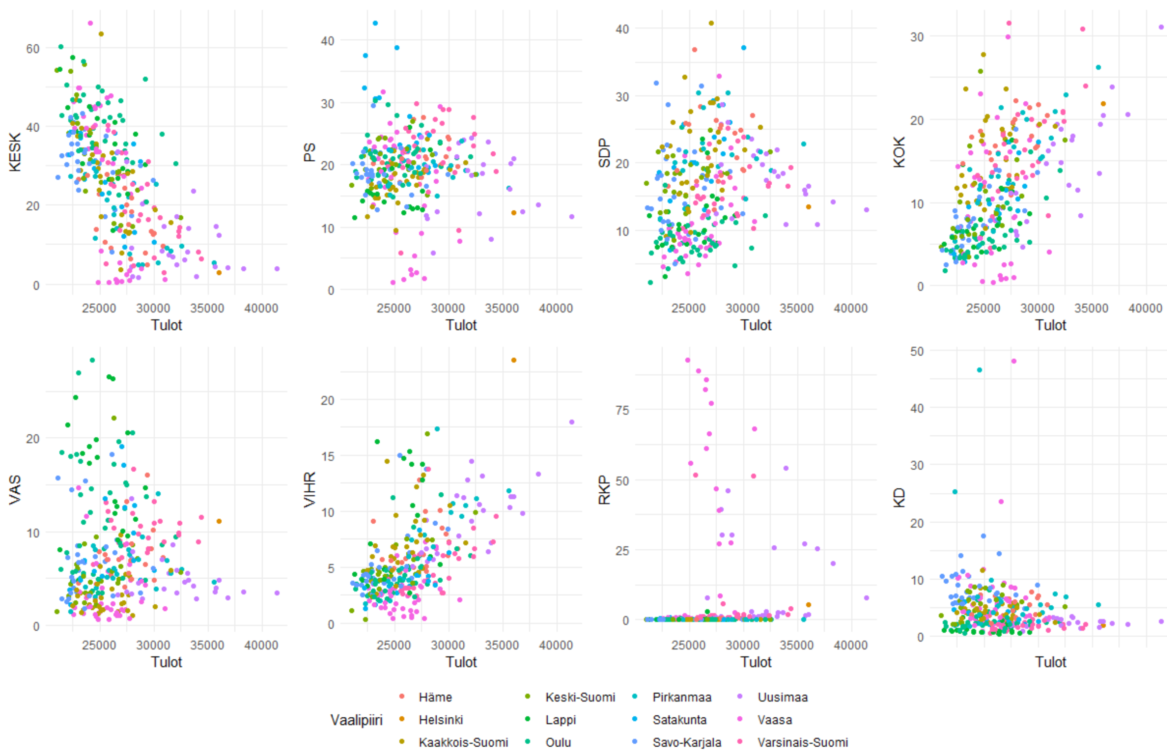


Figure 7: Scatterplot of party support and income in different regions

A cleared scatter plot of Tulot and RKP support is shown in Figure 8.

Figure 8 for the RKP analysis shows that all the observations above around 5% support come from areas in the Vaasa, Uusimaa and Varsinais-Suomi region. The observations from Vaasa particularly show the highest support for RKP. It is likely that these areas have a high proportion of Finnish-Swedish inhabitants, therefore making it reasonable that the RKP party would achieve such high support in those municipalities. Regarding the other parties, some observations could be made about a possible relationship between voting behaviour and main region (Vaalipiiri). Areas from regions with

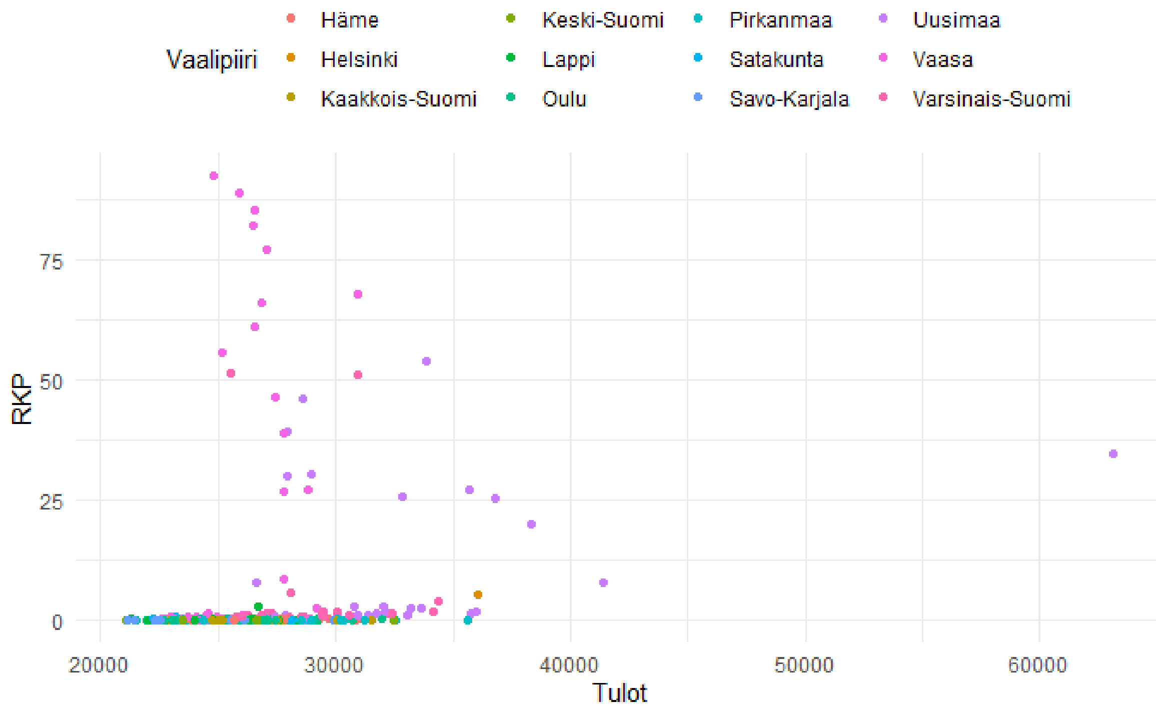


Figure 8: Scatterplot of RKP support and income in different regions

low to moderate income level such as Oulu, Lappi, Savo-Karpela tend to be more represented in the high support areas for KESK, PS and VAS. Areas from higher income regions such as Uusima tend to appear in the higher support range for KOK and VIHR (and RKP). The party KD on average receives the highest support in those municipalities belonging to the Savo-Karjala region with a couple of high support areas from Pirkanmaa and Vaasa. It is likely that this area have a large proportions of Christian communities. All these observations are made by the visual analysis of Figure 7 and might not be accurate.

4 Interpretation of results

In the data analysis section the simple linear regression model, the multiple regression model and the interaction model were examined for some selected party. The parameter results from each model are summarized in Table 18, and the respective adjusted R^2 values are summarized in Table 19.

Model Type	Independent Variables	Independent Variables	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
Simple Linear model		Tulot	−0.279***	−0.013	0.048***	0.110***	−0.003	0.061***	0.074***	−0.022***
		Ansiotulot	−0.300***	−0.011	0.058***	0.109***	0.008	0.065***	0.069**	−0.022**
		Pääomatulot	−0.094	−0.067	−0.133**	0.248***	−0.250***	0.022	0.286**	−0.043
		VerotRate	−5.722***	−0.219	1.282***	2.084***	0.023	1.195***	1.284**	−0.390**
		ValtionveroRate	−6.326***	−0.645**	0.776**	3.409***	−0.908***	1.638***	1.996**	−0.623**
		KunnallisveroRate	−7.374***	0.157	2.464***	1.642***	1.055***	1.321***	0.627	−0.356
Multiple Linear Model	Tulot & VerotRate	Tulot	−0.134***	−0.021	−0.078**	0.112***	−0.027	0.047***	0.118	−0.028
		VerotRate	−3.179***	0.184	2.771***	−0.044	0.533	0.313	−0.951	0.136
	Pääomatulot & KunnallisveroRate	Pääomatulot	−0.595***	−0.067	0.009	0.395***	−0.219***	0.112***	0.371***	−0.073
		Kunnallisverorate	−8.746***	0.003	2.486***	2.552***	0.551*	1.579***	1.482*	−0.525*
	Ansiotulot & Pääomatulot	Ansiotulot	−0.301***	−0.010	0.062***	0.105***	0.014	0.065***	0.063**	−0.021**
		Pääomatulot	0.062	−0.062	−0.165***	0.194***	−0.257***	−0.012	0.253*	−0.032
Interaction Model	AnsiotulotBraket & Pääomatulot	High	−0.076					0.131**		
		Low	−0.142					−0.060		
		Medium-High	0.178					−0.226***		
		Medium-Low	0.005					−0.186***		
	Tulot & VerotBraket	High			−0.064**					
		Low			0.009					
		Medium-High			0.079					
		Medium-Low			0.054					
Note:										*p<0.1; **p<0.05; ***p<0.01

Table 18: Parameter values with significance level for every model

Model Type	Independent Variables	KESK	PS	SDP	KOK	VAS	VIHR	RKP	KD
Simple Linear model	Tulot	0.421	0.002	0.048	0.332	−0.003	0.315	0.024	0.019
	Ansiotulot	0.446	0.001	0.066	0.299	−0.001	0.333	0.018	0.018
	Pääomatulot	−0.002	0.003	0.012	0.064	0.094	−0.002	0.012	0.0002
	VerotRate	0.426	0.001	0.085	0.285	−0.003	0.293	0.016	0.015
	ValtionveroRate	0.228	0.012	0.011	0.337	0.031	0.242	0.017	0.01
	KunnallisveroRate	0.289	−0.003	0.131	0.070	0.040	0.145	−0.002	0.003
Multiple Linear Model	Tulot & VerotRate	0.437	−0.001	0.101	0.329	−0.003	0.315	0.022	0.016
	Pääomatulot & KunnallisveroRate	0.353	−0.0005	0.128	0.215	0.102	0.179	0.018	0.008
	Ansiotulot & Pääomatulot	0.445	0.003	0.087	0.337	0.098	0.331	0.027	0.016
Interaction Model	AnsiotulotBraket & Pääomatulot	0.446					0.248		
	Tulot & VerotBraket			0.147					

Table 19: Adjusted R^2 values for every model

This section does not aim to state that a variable has a numeric effect of a given size on the party support for a given party. The significance of a parameter can only testify that that parameter is different from zero (hence to reject the null hyphothesis of a change in the independent variable to have no effect on the dependent variable) but does not ensure that the effect of the change in the independent variable will have the magnitude of the parameter estimate. Therefore, in this section the values of the parameter will be considered mostly in terms of the sign that they have, in order to say whether a change in the independent variable has a positive or negative effect on the party support. The magnitude will only be considered when comparing it to the significant parameters of other independent variables. The parties will be analyzed one by one and the best model for each will be picked. As a mean of comparison between model, the adjusted R^2 will be used. The reason the adjusted R^2 is used rather than the R^2 is because the former can tell us whether the addition of independent variables in a new model actually improves the fit more than what would be expected by chance. This means that if the adjusted R^2 increases from the simple model to the multiple model, the integration of the new independent variable improves the fit of the model. If it decreases, then the addition of the new estimator improves the model less than what would be estimated by chance.

4.1 KESK

Three types of models were run for KESK: the simple linear models for the six independent variables, three multiple linear models and the interaction model with AnsiotulotBraket and Pääomatulot. In the simple linear models, all independent variables but Pääomatulot achieved significant parameters at the 0.1% significance level. Additionally, all these parameters were negative, which without further analysis might indicate that KESK support decreases with the increase of income (Tulot, Ansiotulot) and taxation rates (VerotRate, ValtionVeroRate and KunnallisveroRate). This claims does of course not consider possible correlations between the independent variables and omitted variable bias. The best adjusted R^2 in the simple linear models is achieved with Ansiotulot (0.446) and the worst one with Pääomatulot (-0.002).

The multiple linear model with Tulot and VerotRate achieves significant parameter for both variables and an adjusted R^2 that is better of both simple linear models (with Tulot and with VerotRate). This means that the increase complexity of the model actually improves the fit and the multiple linear model is to be preferred to the individual simple linear models of Tulot and VerotRate. It was analyzed in subsection 3.3 that this might occurs due to the fact that the simple linear models suffered from OVB that is at least partially rectified in the multiple model.

The same observations can be made for the multiple linear model with Pääomatulot and Kunnallisvero rate. In this case the adjusted R^2 does not only increase but the parameter estimate for Pääomatulot becomes significant in the multiple model. This was also explained in subsection 3.3. For the third multple linear model, the adjusted R^2 worsens by 0.001 compared to the simple model with Ansiotulot. Additionally the parameter for Pääomatulot remains insignificant and turns positive. The reason for this was tried to be explained using the interaction model with AnsiotulotBraket and Pääomatulot. It was observed that none of the parameter for the Ansiotulot brackets are significant, and that the interaction model did not necessary improve the fit (the adjusted R^2 remains equal to the one achieved with the simple linear model with Ansiotulot).

Given these observations, it might be concluded that the simple model with independent variable Ansiotulot explains the party support for KESK the best. That is, party support decreases with an increase in earned income. We have observed that the multiple linear model with Tulot and VelotRate also achieves a quite good adjusted R^2 that improves from both individual linear model, which means that both Tulot and VerotRate seem to have an individual effect on the party support. Nevertheless in order to claim that we should first check the correlation that these two parties have with Ansiotulot, as it could be the cause behind the improvement of the adjusted R^2 since both variables are correlated to it.

4.2 PS

The support for PS stands out from our models for having the least amount of significant parameters. The only significant parameter in the simple models is achieved with the independent variable ValtioveroRate. The parameter is also negative, which might indicate that the party support decreases with an increase in the state taxes. Nevertheless, the adjusted R^2 is very poor (0.012) despite being the best one among all models for PS.

To be fair, the apparition of a possible relationship between ValtionveroRate and party support could be due to chance in this case. ValtioveroRate is strongly correlated to Tulot, yet the parameters for Tulot in the simple model and multiple model are insignificant. Another plausible reason for the negative relationship between ValtioveroRate and party support might be, for example, that the party wants to instigate some policies that might worsen the situation for people having higher taxation rates. A not so educated guess could be making the margins for progressive taxation stricter and smaller.

4.3 SDP

SDP is the party for which the increase complexity in the multiple models seem to improve the fit, despite the fact that adjusted R^2 values remain comparably small when put side by side with the other parties. Nevertheless the highest adjusted R^2 is achieved with the simple model with KunnallisveroRate (0.131). The R^2 worsens slightly in the multiple model with Pääomatulot and KunnallisveroRate, indicating that in this case the increased complexity does not add anything to the fit. The parameter value is also very high (2.464), meaning that at a 1% increase in the KunnallisveroRate there should result in a 2.46% increase in the party support. Politically, a hypothetical reason this might be is related to possible municipality tax reductions promised by the party. If we also consider the interaction model with Tulot and VerotBraket, the R^2 is actually the highest in the interaction model (0.147) but only the parameter for the "High" VerotBraket is significant. The reason for this was considered in subsection 3.4 and no definite answer was found. All of this consider, it is probably best not to claim any clear relationship with variables and SDP support, since the adjusted R^2 is poor in every model, despite the fact that many parameters result significant. The most acceptable claim would be that there seem to be an increase in support as average income increases.

4.4 KOK

For KOK, the highest adjusted R^2 values are achieved in the simple model with ValtionveroRate and in the multiple model with Ansiotulot and Pääomatulot. These are both 0.337. All parameters in the simple model are positive. The R^2 does not improve for the multiple model with Tulot and VerotRate, but it improves for the multiple model with Pääomatulot and KunnallisveroRate and the multiple model with Ansiotulot and Pääomatulot, indicating that the added complexity improves the fit. In other words, the behaviour for KOK support is better explained by considering these variables together than separated, enhancing that each variable has their individual effect on the support. Additionally, all the parameters in the multiple models are significant. In hindsight, it could have been useful to run a multiple linear regression for KOK with Ansiotulot, Pääomatulot, KunnallisveroRate and ValtionveroRate as estimator. Out of curiosity, the regression was run and can be found in Appendix 7.2. In the new multiple regression, the adjusted R^2 improves to 0.383 but KunnallisveroRate and Ansiotulot lose significance, which might indicate that actually the best model for KOK might be a multiple model with Pääomatulot and ValtionveroRate. This model was also run and the result are shown in Appendix 7.3. The adjusted R^2 was the highest for KOK so far, 0.387, and both estimator achieved significant parameter values at the 0.01% level. This said, it might be quite safe to say that among all the model run for KOK, the multiple linear model with estimators Pääomatulot and ValtionveroRate is the best fitting. According to this model, KOK support decreases as Pääomatulot increases and increases as ValtionveroRate decreases. The magnitude for the ValtionveroRate variable is much higher (4.77).

4.5 VAS

VAS shows the highest adjusted R^2 in the multiple model with Pääomatulot and Kunnallisvero. Nevertheless the fit is still quite poor (0.102). It possible that a neater explanation for VAS support stands on the analysis made in subsection 3.5. One thing that is possibly relevant is that VAS support seems to decrease with the increase in Pääomatulot more than any other party except RKP. Another wild guess might be that this relationship might happen because VAS aims to for example higher the tax on investment income.

4.6 VIHR

The best R^2 for VIHR is achieved in the simple model with Ansiotulot. Most of the variables show a significant parameter in the simple models except Pääomatulot, but as for every other party this might actually be caused by the correlation between the independent variables that results in significant parameters even though the independent variable itself might actually not be the one causing a relationship with the support (OVB). The multiple model with Pääomatulot and KunnallisveroRate shows an adjusted R^2 that is significantly better than the simple models with Pääomatulot and KunnallisveroRate, yet the best adjusted R^2 still remains for the simple model with Ansiotulot. The interaction model was analyzed more thoroughly in subsection 3.4.

4.7 RKP

The RKP case was already brieftly discussed previously, at it was enhanced the idea that the support might rely more on the region (Vaalipiiri) than the income variables or taxation variables themselves, mostly because the RKP party targets a specific part of the finnish population, that is the Finnish-Swedish people. The models return adjusted R^2 values almost as poor as those for PS, but this is likely the result of those outliers that relevantly increase the standard error and therefore worsen the adjusted R^2. In this case, it could have probably been more useful to analyze the model with and without the RKP support outliers. The highest adjusted R^2 is achieved in the multiple linear model with Ansiotulot and Pääomatulot, even though the parameter for Pääomatulot is not significant. Nevertheless from the significance of Pääomatulot and Ansiotulot and the aggregate Tulot in the simple models, it could be said that if it weren't for those wild outliers RKP would have shown more clearly an increase in support with an increase in income variables.

4.8 KD

KD shows overall very poor results for the adjusted R^2 for all models. The simple linear models show significant parameters for most variables except Pääomatulot and KunnallisveroRate. The highest adjusted R^2 is achieved in the simple model with Tulot and the none of the multiple model improves the adjusted R^2 from the simple models, indicating that the additional complexity does not improve the fit.

Nevertheless, the adjusted R^2 for the simple model with Tulot is only 0.019 and the magnitude of the parameter is very small (the parameter value is -0.022).

In conclusion, KD support appears to decrease with the increase in income, but the fit of the model is poor.

5 Improvements

Different approaches could have been used to analyze the data. Linear regression was chosen because it did not seem logically reasonable to use higher order exponents for the independent variables, but this was just assumed and no analysis was conducted on higher order regression polynomials. It is possible that many of the models run are highly biased, especially the simple models. That occurs because the conditions $cov(X, \epsilon) = 0$ for an unbiased least-square parameter might be violated. This means that the error term is correlated with the independent variable(s) and the model suffers from OVB. It was tried to rectify the change of OVB in the multiple models by adding estimators that showed correlations with each other to see whether the individual independent variables had a significant effect on the party support. More attention could been paid towards finding instrumental variables that would limit the chance of OVB. To clarify, an instrumental variable is a variable correlated to the independent variable but uncorrelated to the error term. This instrumental variable has an effect on the dependent variable only thanks to the correlation that it shows with the independent variable.The multiple models conducted in subsection 3.3 tried to introduce some of these instrumental variables in the model. For example, VerotRate was added to the model with Tulot because it showed high correlation with Tulot and most of party variables (Figure 3). Nevertheless instrumental variables are difficult to recognize and further analysis should have been conducted to analyze the multiple linear regressions under the instrumental variable analysis point of view.

The adjusted R^2 was used to compare models for the reasons previously explained in section 4. Other methods of comparisons between model could have also been used in addition or in place of the adjusted R-squared. These could have been the AIC, BIC and the F-statistics p-value, among others such as the residual standard error.

6 Conclusion

As mentioned in the introduction, it was tried to give some explanation for the support of each of the eight chosen parties. It was seen that for most parties the support is likely best explained by only one explanatory variable, that more often than not used to be the variable for the average earned income Ansiotulot. Some party support was better explained by multiple variables, for example in the case of KOK. For some parties the support was more dependent on the region of the municipalities, for example RKP. The writer had very little background knowledge about the Finnish parties analyzed. This was probably a good thing in the sense that the data analysis was not biased towards selecting models that would align with the parties aims and policies. On the other hand, some starting idea could have been helpful. They will absolutely inform themselves about the parties as soon as possible just to see whether some of their guesses had some true grounds.

A Appendix

A.1 Scatter plots for simple linear regressions

Figure 9: Party support and average earned income scatter plots

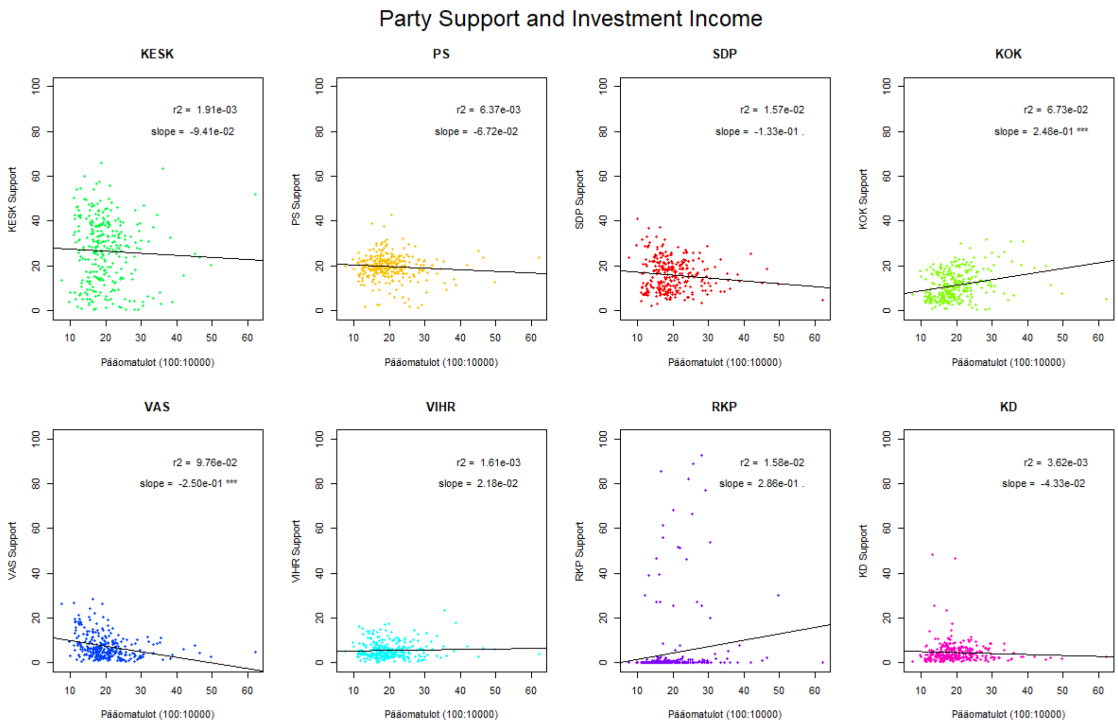


Figure 10: Party support and average investment income scatterplots

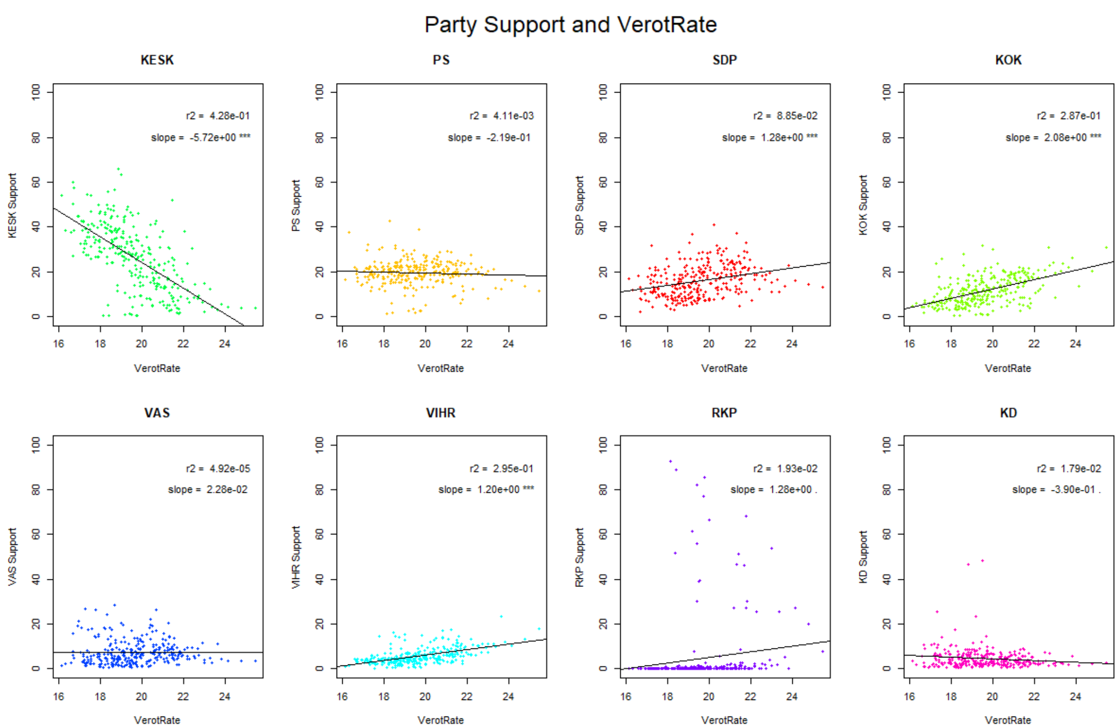


Figure 11: Party support and tax rate estimate scatterplots

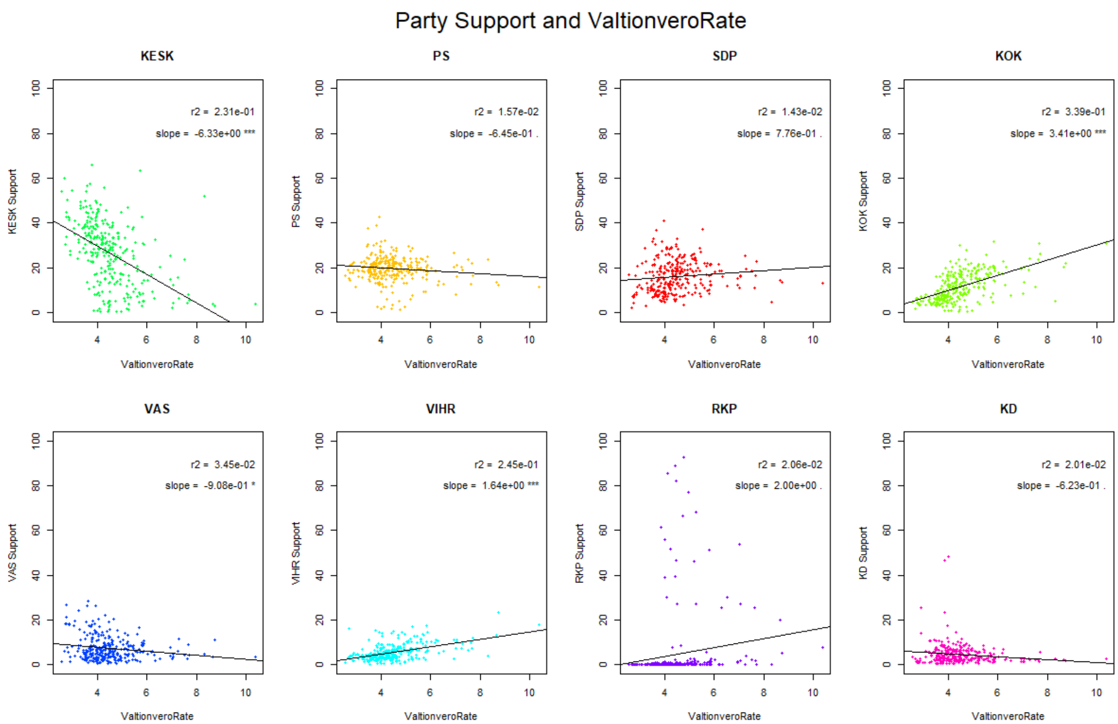


Figure 12: Party support and state tax rate estimate scatterplots

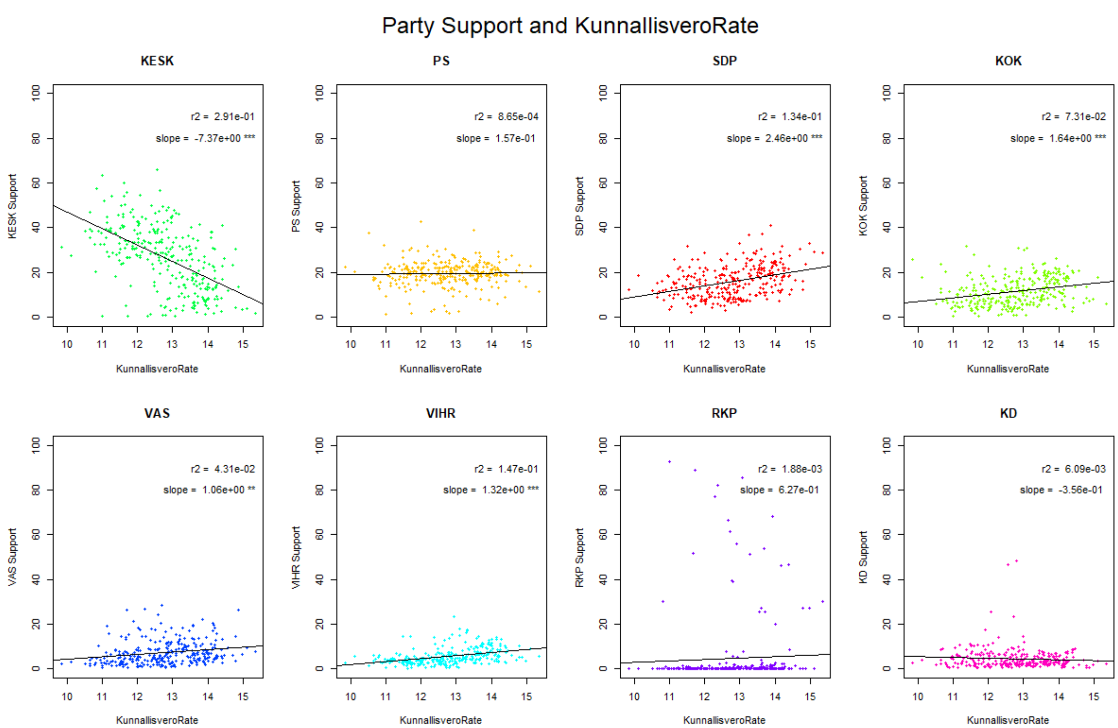


Figure 13: Party support and municipality tax rate scatterplots

A.2 KOK multiple regression with three variables

	Dependent variable:
	KOK
Ansiotulot/100	−0.007
	(0.031)
Pääomatulot/100	-0.343^{***}
	(0.120)
ValtionveroRate	5.072^{***}
	(1.078)
KunnallisveroRate	−0.003
	(0.522)
Constant	−2.486
	(5.260)
Observations	294
R^2	0.391
Adjusted R^2	0.383
Residual Std. Error	5.031 (df = 289)
F Statistic	46.438^{***} (df = 4; 289)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 20

A.3 KOK multiple linear regression with Paaomatulot and ValtionveroRate

	Dependent variable:
	KOK
Pääomatulot/100	-0.312^{***}
	(0.063)
ValtionveroRate	4.777^{***}
	(0.384)
Constant	-3.652^{***}
	(1.235)
Observations	294
R^2	0.391
Adjusted R^2	0.387
Residual Std. Error	5.015 (df = 291)
F Statistic	93.442^{***} (df = 2; 291)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 21