Explaining the non-voters in Finnish 2007 parliamentary elections

Matti Nelimarkka*

April 11, 2010

Abstract

In this study voting behaviour is examined. The data used is from a survey conducted after 2007 elections, having total of 1422 responders. Of these, 1218 cases are valid and forms the sample, where the population is all persons entitled to vote in the 2007 election.

We examine sex, age, income and education and observe, that age and education are useful predictors. We also noted that the income variable has an effect, but it is not clear if there exists an interaction between age and income.

Lastly, we discuss if the background questions, which were used here, build a good prediction model. The overall correctness in the final model is 85.2 %, but of the non-votes, less than 10 % are predicted correctly. Thus, we argue that other variables, which may be related to values and believes should be used in the further research to allow better estimates.

* Student ID: 013298692
matti.nelimarkka@helsinki.fi

Contents

1 Introduction **4**

2 Previous research **4**

3 Data **6**

4 Analysis **10**

4.1 The method 10

4.2 M_A: The age 11

4.3 M_B: The education 12

4.4 M_C: The income 13

4.5 M_D and M_E: Interaction with age, sex and income 14

5 Discussion **15**

6 Conclusion **20**

List of Figures

1 The distribution of responders age ($n = 1218$) 8

2 The distribution of responder's household's income ($n = 1218$) 9

3 The effect of age to voting 18

4 The effect of income to voting 19

5 The effect of income and age to voting 19

List of Tables

1 Explanatory variables in Finnish research 6

2 Demographics of the data ($n = 1218$) 6

3 Item missing per variable ($n = 1422$) 7

4 Educational level of the sample ($n = 1218$) 7

5 Examination of the relationship between the sex and voting . 12

6 The effect of age controlled with sex 12

7 Examination of the relationship between the education and voting 13

8 The effect of education 13

9 The effect of income 14

10 The interaction of income with age 15

11 The interaction of income with sex and age 16

12 Summary of examined models 16

13 The final model 17

14 The elaborated final model 17

1 Introduction

Lijphart's (1997) article *'Unequal Participation: Democracy's Unresolved Dilemma'* points out that in an ideal democracy the participation should be equal. However, the research has indicated that participation in modern societies is far from equal. For example, those who have higher social class are more likely to participate than those with lower social class. Even though it is said that voting is more equal compared to other methods of participation, it too has some drawbacks[1]. He later observes, that the topic of unequal participation is even more relevant in the modern era, as the voting turnout has decreased over time (see e.g. Wattenberg 2000). Thus, the initially unequal participation has become even more unequal due to the changes in participation.

However, it is relevant to understand that voting is only one form of participation. The topics studied in political participation vary a lot (see e.g. van Deth 2001), but this study uses the most traditional concept of voting as the key form of participation. This is due to first data access – study of voting behaviour has been done for a longer time period and it is easier to quantify – and secondly as participation via voting is still the *de facto* method of participation as we life in a representative system.

In this study we examine the Finnish political context based on 2007 elections. The article is structured as follows. First, we examine the previous research on voting behaviour. We also define the hypothesis (see page 5) that we shall study. Then, we describe the data and the methods used in this study. After this, we provide analysis and discussion of the topic. Lastly, we conclude our results.

2 Previous research

As discussed above, there has been studies related to voting behaviour. Topf (1995) discusses variety of this research and indicates, that in general, education and sex are not significant factors, where as age, especially when seen as generations, is more explanatory.

[1]And some of these drawbacks are handled via institutional design methods, such as compulsory voting. Even while this is interesting, this work here is not planed to suggest or examine institutional settings but just examine situation in Finland.

More relevant ones are however those where the political context is more similar to Finnish environment. For example Wass (2007) observed, that younger people vote less actively than older people. What is special in Wass's (2007) study is the longitudinal nature: it studies parliamentary elections since 1979 and observes that there is a clear generation shift, i.e. younger generations in general are less likely to participate than older ones – not just younger people. In this study however, as we only work with single snapshot, the observation of linearity in age and turnout.

Martikainen, Martikainen & Wass (2005) note, that the income, home tenure and education have a strong effect on voting behaviour. They report, that the education has a stronger effect with younger persons where as income has a stronger effect with older persons. For people under 25 higher education is a good predictor of turnout and for people over 25 higher income and housing tenure is good predictor of income. Lastly, they also observed different social classes and found out that higher social classes are more active, farmers being an exception.

Based on the previous research shown above, we shall examine the following hypothesis

- H_1 Younger responders are less likely to vote (Wass 2007).
- H_2 Less educated are less likely to vote (Martikainen et al. 2005).
- H_3 Persons with lower income are less likely to vote (Martikainen et al. 2005).
- H_4 There is an interaction between the age and the income (Martikainen et al. 2005).

This leads us to only use four explanatory variables[2], which is not high number considering the phenomena. In table 1, we examine Finnish research concluded in this area, which indicates that four variables is not especially low compared to other works. Compared to that research, the only missing variable is related to social class. The social class is asked in the survey (question K5), but as it is self-perceived, the use of it is harder compared to the four variables[3].

[2]Which are sex, age, education and income, further discussed in the next section.

[3]For example, Martikainen et al.'s (2005) and Wass's (2007) operationalization is to five categories: farmers, entrepreneurs, upper middle class, lower middle class and manual workers where as in this data, working class, lower middle class, middle class, upper middle class and upper class is used.

Research	Explanatory variables	Total
Wass (2007)	Sex, social class, age, generation, period	5
Martikainen et al. (2005)	Age, sex, social class, education, income, housing tenure	6
Martikainen & Wass (2007)	Age, sex, location	3

Table 1: Explanatory variables in Finnish research
Observe that Martikainen & Wass (2007) is not published in a peer reviewed journal.

Voted in 2007 elections	Yes 1026 (84.2 %)	No 192 (15.8 %)
Sex	Male 626 (51.4 %)	Female 592 (48.6 %)
Age	$\overline{x} = 49.2939$	$\sigma = 17.82898$

Table 2: Demographics of the data ($n = 1218$)

3 Data

This study uses the '*Finnish National Election Study 2007*' (FSD 2269) from the Finnish Social Science Data Archive[4], collected by Taloustutkimus Oy.The study was conducted after the parliamentary elections in 2007 with first an interview and after that a postal survey. The population is all persons entitled to vote in the 2007 elections, excluding Åland. The sample was selected using quota sampling based on age, gender and province of resistance. The total number of responders was 1422, however in this study $n = 1218$ due to missing data (for details, see Table 3). The data file does not include information of response rate nor the cross-sample size, but the quota method should allow representative selection of citizens being asked to vote[5].

Some of the democratic are presented in Table 2. Clearly there is majority of sample have voted, which is not an surprise, considering that majority (67.9 %) of Finnish people voted. However, this sample indicates a significantly higher turnout of 84.2 %. This kind of over reporting is well known in the field (eg. Silver, Anderson & Abramson 1986, Bernsteind, Chadha & Montjoy 2001). It is indicated, that those who are highly educated, have

[4]The author is responsible the analysis and results presented here.

[5]Actually, there was a sample weight used as there are more Swedish speaking in this sample compared to population. In this study, the weights are **not** used due to author's mistake. However, even with the weightings in use, there are certain issues, such as the age distribution.

	Valid	Missing	(%)
Voted in 2004 elections	1420	2	0.1
Sex	1422	0	0.0
Age	1422	0	0.0
Education	1422	0	0.0
Income	1219	203	14.22

Table 3: Item missing per variable ($n = 1422$)

		Basic education				
		Elementary	Upper elementary	Still in school	Upper secondary	None
Vocational education	Still in school	4	19	4	14	0
	Secondary school	105	243	4	20	1
	College degree	24	155	2	90	1
	Some uni. or poly. studies	2	16	0	64	0
	Polytechnic degree	1	26	2	56	0
	University degree	3	6	0	173	0
	None	74	66	6	37	0

Table 4: Educational level of the sample ($n = 1218$)

good income etc. report more likely that they have voted when they actually have not. Thus, this is an obvious source of bias, but can not be controlled by the researcher[6].

Secondly, we observe a slight over representation of males. In the initial data set there was a bit more of females (720 vs. 702). While filtering, one of the criteria was an valid household income (D20), where females had more invalid values (127 vs. 76), which is one core reasons to this kind of difference (also, it should be observed that this variable is the most item missing variable, shown in Table 3).

The age of the responders varies from 17 to 94. The arithmetical mean is about 49 years. However, we observe that the distribution has strong kurtosis even while the skewness is not high (see Figure 1). The effect of this to the analysis is not clear, however having a strong kurtosis means that the results we got from the binary logistic analysis may be wrong.

The education-variable was asked in two questions (D3.1 and D3.2). The first question was related to basic education and the second one to vocational education. The first one included options about elementary school education level where as the second one was related to university education level (see Table 4). The variable was first categorized in three section

[6]Also, the usual problems of survey methods, i.e. the bias of those responding to surveys, is effecting in this problem and also may lead to other biases when compared to population.

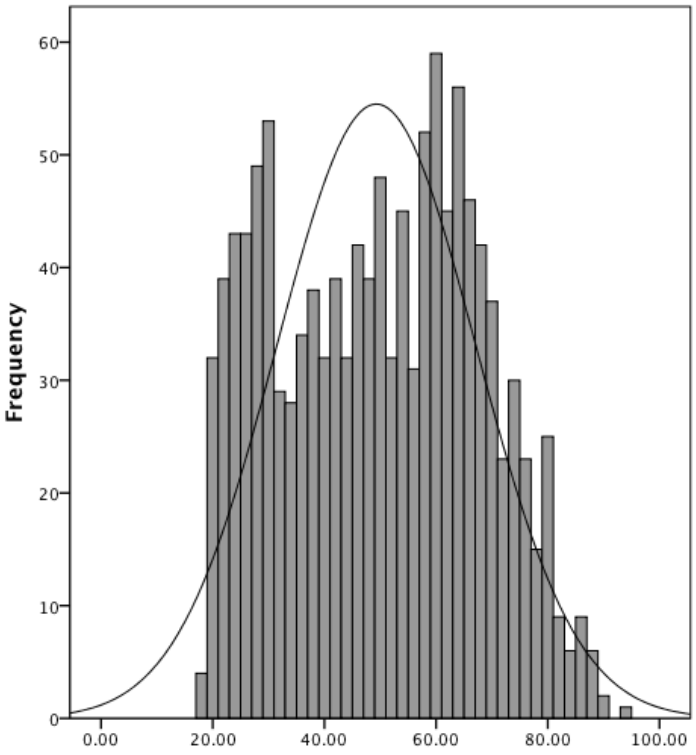

Figure 1: The distribution of responders age ($n = 1218$)

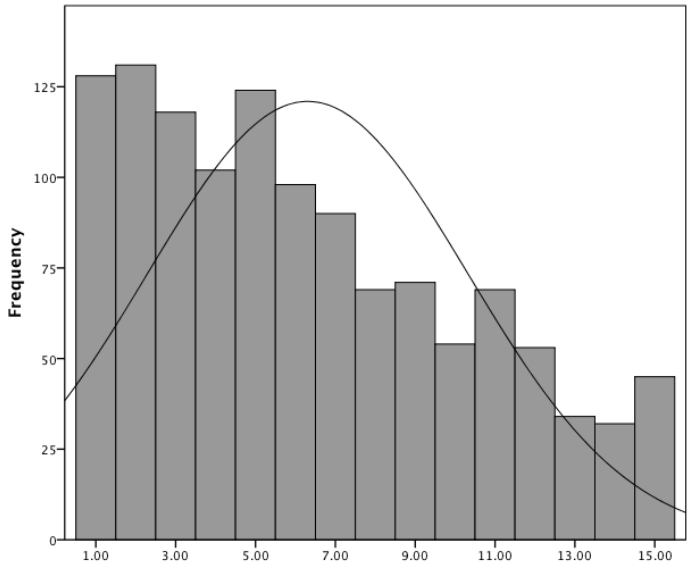

Figure 2: The distribution of responder's household's income ($n = 1218$)

1. Having elementary or upper elementary level education
2. Having secondary or upper secondary level education
3. Having university or polytechnic level education

Lastly, we examine the income as a dependable variable. The survey requested responder to estimate household's yearly income (D20) as a continues variable. The data only included a discrete variable of income, which started from €10 000 or less. The variable was dived in €5 000 categories until €90 000 or more, which was the last category. Even when the variable is a categorical, the author uses it as a continuous variable. This is mainly due to the fact, that variable has 15 categories. There are however two problems in this variable considering analysis.

First, as one can observe from the Figure 2, there is negative skewness and high kurtosis. Thus, as with age, the binary logistic regression does not recommend use of this kind of variable. Again, this may be reflected in the results of the analysis. Secondly, the question is related to house hold income, where as H_3 deals with person's income. Thus, this operationalization of variable is not the best considering the theory, but as the data does not include personal income this is the only method to include this kind of variable. This effect could be controlled using the knowledge of the household

size (D21.1) to give a better estimate, but this indicator also has problems, such as the handling of minors[7] or the different income of persons in the household, and is thus not conducted in this study.

We have now examined the variables and described the possible issues in them. We will in the next section engage analysis of different hypothesis presented above. After this, we shall discuss the results and provide conclusions.

4 Analysis

We analyse the four hypothesis set in page 5. Those were related to the effect of age, education and income in the voting behaviour. The last hypothesis was related to the interaction of both age and sex to the income. However, we first examine the method and the measures used in more detail.

4.1 The method

In the ordinary least squares regression (*OLS regression*) the coefficients $\beta_0, \beta_1, \beta_2 \ldots \beta_n$ of variables $x_1, x_2 \ldots_n$ in equation $y = \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \beta_0 + \varepsilon$ are calculated so, that the error term, ε is lowest (Lewis-Beck 1980, 47–51). However, let us assume that dependent variable, y has binary nature of, i.e. it only has two values in the data. The OLS regression can not handle the natural restriction that $y \in [0, 1]$, and thus a new model of calculating, *binary logistic regression* is needed (Pampel 2000, 1–5). Obviously in this case, when we explain if person voted or did not vote, the dependent variable has only binary values and thus requires the use of binary logistic regression, here after discussed as regression.

However, the logistic regression rises some issues. The nature of this model, rather than linear is logarithmic. The effect of this may be observed e.g. in Figure 3 (on page 18). Observe the curve-nature of the effect plot. Thus, the effect of regression coefficient β_n is not linear but the effect can be shown as odds-ratio (for details, refer Pampel 2000, 21–23), which can be interpreted as the effect of change of one unit, similarly to OLS regression. The calculation is based on formula $(e^{\beta_n} - 1) \cdot 100\%$, but in this work, the calculation is done by PASW Statistic version 18.0 (former SPSS) (Pampel 2000, 5–39).

[7] Questions D21.2, D22.1 and D22.2 include data regarding the number of different age segments, but not the income of persons.

The model wellness can be also calculated. First measurement is the Log Likelihood–measurement, here called the Omnibus test, which tests if the regression coefficients $\beta_0, \beta_1, \beta_2 \ldots \beta_n, \beta_{n+1} = 0$. Here the interpretation is done especially when adding new variables, x_{n+1} to the equation. If the Omnibus test indicates, that the hypothesis of coefficients, especially the change due to newly added β_{n+1}, being null is not true (the value of the χ^2-based test is less than 5 %), adding a new variable to the equation allows better estimation (Menard 2002, 20–24).

Secondly, Hosmer–Lemeshow test the distribution of the estimated probabilities by grouping it into certain groups. Without more closer focus to details, which are said to be concluded based on simulations reported in their earlier work, the logistic regression is '*correct model*' when the Hosmer–Lemeshow goodness-of-fit statistic is similar to χ^2 distribution with number of groups minus two degrees of freedom. The use of 10 % percentiles gives the best results and thus is compared with χ^2 distribution of 8 degrees of freedom (Hosmer & Lemeshow 2000, 147–156). As the we want the distribution being similar, in this case we prefer that $p > 0.05$, which indicates similarity, where as $p \leq 0.05$ would lead to the acceptance of difference.

Lastly, we often discuss the changes in the predictive efficiency, i.e. in how many cases the model fails. There are several methods to compare this, such as the changes of error in the estimation caused by the model or calculations derivative from the 2×2-table of observed and predicted values (Menard 2002, 27–36). In this work, we report often the overall correctness of the model and the changes in the predictions caused by different models, but do not use the more advance methods suggested in Menard's (2002) work.

4.2 M_A: The age

First examine the effect of age as an estimate when we control the sex. Based on the previous research, we do not assume relationship between sex and voting behaviour. This data also suggests, that there's no differences between sexes ($p = 0.601$), as show on Table 5. On the other hand, there is a positive, not strong, but existing positive correlation ($r = 0.234$, significant in 1 %–level) between the age and the voting in 2004 elections. Thus, this initially promises support that age should be included in the analysis.

The first model is not promising in any way. Firstly, the model estimates that everyone would have voted, meaning in practice 84.2 % correctness. The

	Did not vote	Voted	$\sum$
Female	90 (7.4 %)	502 (41.2 %)	592 (48.6 %)
Male	102 (8.4 %)	524 (43.0 %)	626 (51.4 %)
$\sum$	192 (15.8 %)	1026 (84.2 %)	1218 (100 %)

Table 5: Examination of the relationship between the sex and voting
$\chi^2 = 0.273, df = 1; p = 0.601$

Factor	β	Significance
Sex	-0.115	0.481
Age	0.101	<0.001
Age^2	-0.001	0.015
Constant	-1.260	0.024

Table 6: The effect of age controlled with sex
Sex: Male = 1.

model does however have Hosmer and Lemeshow level test in the valid area, i.e. $p = 0.614, \chi^2 = 6.298, df = 8$.

If we examine the regression equation (see Table 6) only one factor is statistically significant (the age). However, this is also the factor which was assumed to have a positive effect, as the H_1 states: the older people are more likely to vote. In the next step we add the educational dummies to control the effect.

4.3 M_B: The education

The education was categorised in three classes (see above, page 3), the low level of education (only primary or upper primary), the middle level (secondary or upper secondary education) and higher level of education (university or polytechnic degree). Thus, there are two dummy values, one indicating the lower level of education and one indicating the higher level. Based on H_2 we assume, that those with higher level of education are more likely to vote and those with lower level of education are less likely to vote. The univariate examination indicates support for difference based on education ($p < 0.001$, see Table 7).

This hypothesis is supported by the data with the control of age. As Table 8 indicates, those with higher education are four times more likely to vote compared to those with medium level of education. On the other hand,

	Did not vote	Voted	$\sum$
Low level	43 (3.5 %)	130 (10.7 %)	173 (14.2 %)
Middle level	136 (11.2 %)	642 (52.7 %)	778 (63.9 %)
High level	12 (1.1 %)	254 (20.9 %)	267 (21.9 %)
$\sum$	192 (15.8 %)	1026 (84.2 %)	1218 (100 %)

Table 7: Examination of the relationship between the education and voting $\chi^2 = 36.364, df = 2; p < 0.001$

Factor	β	Significance	Odds ratio
Sex	-0.057	0.734	0.945
Age	0.070	0.012	1.072
Age^2	0.000	0.306	1.000
Low education	-0.712	0.001	0.491
High education	1.457	<0.001	4.293
Constant	-0.841	0.152	–

Table 8: The effect of education
Sex: Male = 1. Low education: primary level of education as highest level. High education: university or polytechnic as highest level of education.

likelihood of those with low level of education is approximately half of those with medium level of education. Both of these are significant in the 1 %-level.

What is the effect of this addition to the model in general. Firstly, the model now estimates 25 persons not voting (and making mistake in 10 cases), which means that the overall correctness is slightly up to 84.6 %. The Hosmer–Lemeshow test indicates also validity ($p = 0.384, \chi^2 = 8.525, df = 8$). Also, the Omnibus test indicates, that the dummy variables should be kept in the model, as the test is significant in 1 %-level ($\chi^2 = 53.146, df = 2$). Thus, the conclusion is, that education should be kept as an factor.

4.4 M_C: The income

Now, the education controlled with the income-variable. Even though the variable, as discussed above, has certain problems, the income hypothesis that those with high income vote more likely is valid based on this data. The univariate examinations shows a weak positive correlation ($r_p = 0.154$, significant in 1 %–level). The income is significant in the 5 %-level even when controlled for sex, age and education (see Table 9).

Factor	β	Significance	Odds ratio
Sex	-0.110	0.516	0.896
Age	0.043	0.148	1.044
Age^2	0.000	0.947	1.000
Low education	-0.650	0.004	0.522
High education	1.260	<0.001	3.526
Income	0.073	0.004	1.075
Constant	-0.642	0.282	–

Table 9: The effect of income
Sex: Male = 1. Low education: primary level of education as highest level. High education: university or polytechnic as highest level of education.

We also observe from Table 9, controlling of the income decreases the odds of higher education and increases it for those with lower education – in practice meaning that controlling income decreases the importance of education.

The model now estimates 85.1 % of the cases correctly, major improvement has been on the decrease of those estimated not voted but still said to vote (from 10 to 7). Still, however the model can only predict 8,9 % correct of those not voted. The Hosmer-Lemeshow test is till good ($p = 0.642$) and there has been decrease in the χ^2 to 6.043 ($df = 8$). The omnibus test indicates, that this variable can be added, as it is statically significant in 5 %-level ($\chi^2 = 8.495, df = 1$).

4.5 M_D and M_E: Interaction with age, sex and income

We lastly examine the interactions of age, sex and income. We first add interaction of age and income, resulting model M_D and then examine the addition of age in the model M_E.

The results of the first interaction are shown in Table 10. The Omnibus test does not support adding this interaction, as the significance is above 5 %-cut-point ($p = 0.074, \chi^2 = 5.217, df = 2$). Also, the over all model fit is lower ($\chi^2 = 11.030, df?8, p = 0.200$), but still acceptable.

Lastly, we examine the interaction of both age and sex to the income. The Table 11 shows interaction of both age and sex. The omnibus test indicates that the addition of these variables is not recommended ($p = 0.307, \chi^2 = 1.044, pdf = 1$). However, the Hosmer and Lemeshow test is still valid,

Factor	β	Significance	Odds ratio
Sex	-0.139	0.412	0.870
Age	0.063	0.160	1.065
Age^2	0.000	0.408	1.00
Low education	-0.622	0.005	0.537
High education	1.285	<0.001	3.615
Income	0.165	0.370	1.179
Income and age	0.008	0.371	0.992
Income and age	0.000	0.210	1.000
Constant	-0.761	0.480	–

Table 10: The interaction of income with age
Sex: Male = 1. Low education: primary level of education as highest level. High education: university or polytechnic as highest level of education.

$p = 0.746$ ($\chi^2 = 5.109, df = 8$).

5 Discussion

We have above tested five different models (see Table 12) to explain why some people vote and others do not vote. As the final model, we vary the model C. First, it is the best model considering overall correctness. Secondly, the omnibus test indicate, that in models D and E, the addition of these variables is not recommended. Third, the Hosmer and Lemeshow test is well above the 5 %-cut point. We call the varied model as F.

From the variables in model C, which were sex, age, age^2, education and income, we note, that neither sex nor age (and age^2) are statistically significant. Thus, the examination would suggest dropping of these variables. Considering the sex, the previous research in Finland has not discussed is as a theoretically relevant. But, in the case of age, previous research supports the existence of this relationship. We also note, that the curve-linear interaction term is less significant ($p = 0.947$) and the coefficient of this variable is 0.000, meaning that the accuracy of the software is not good enough to calculate the coefficient. Thus, we examine as the first final model, F_1, the model without the curve-linear age and the sex.

The results of the first final model are presented in Table 13. The Hosmer and Lemeshow Test indicates, that the model is good, $p = 0.737$ ($\chi^2 =$

Factor	β	Significance	Odds ratio
Sex	-0.384	0.216	0.681
Age	0.065	0.143	1.068
Age^2	0.000	0.373	1.000
Low education	-0.624	0.005	0.536
High education	1.295	<0.001	3.652
Income	0.147	0.349	1.158
Income and age	0.008	0.349	0.992
Income and age^2	0.000	0.197	1.000
Income and sex	0.048	0.306	1.049
Constant	-0.687	0.457	–

Table 11: The interaction of income with sex and age
Sex: Male = 1. Low education: primary level of education as highest level. High education: university or polytechnic as highest level of education.

Model	Table	Variables	Hosmer and Lemeshow	Overall correctness (%)
A	6	Sex, age, age^2	$p = 0.614, \chi^2 = 6.298, df = 8$	84.2
B	8	Sex, age, age^2, education	$p = 0.369, \chi^2 = 8.687, df = 8$	84.6
C	9	Sex, age, age^2, education, income	$p = 0.642, \chi^2 = 6.043, df = 8$	85.1
D	10	Sex, age, age^2, education, interaction of income and age	$p = 0.200, \chi^2 = 11.030, df = 8$	85.0
E	11	Sex, age, age^2, education, interaction of income and age, interaction of income and sex	$p = 0.746, \chi^2 = 5.109, df = 8$	84.8
F_1	13	Age, education, income	$p = 0.737, \chi^2 = 5.187, df = 8$	85.2
F_2	14	Age, education, income, interaction of income and age	$p = 0.933, \chi^2 = 3.018, df = 8$	84.9

Table 12: Summary of examined models

Factor	β	Significance	Odds ratio
Age	0.041	<0.001	1.042
Low education	-0.655	0.003	0.519
High education	1.273	<0.001	3.571
Income	0.071	0.003	1.074
Constant	-0.656	0.009	–

Table 13: The final model
Sex: Male = 1. Low education: primary level of education as highest level. High education: university or polytechnic as highest level of education.

Factor	β	Significance	Odds ratio
Age	0.029	<0.001	1.029
Low education	-0.623	0.005	0.536
High education	1.287	<0.001	3.622
Income	-0.047	0.476	0.954
Income and age	0.003	0.068	1.003
Constant	-0.182	0.609	–

Table 14: The elaborated final model
Sex: Male = 1. Low education: primary level of education as highest level. High education: university or polytechnic as highest level of education.

$5.187, df = 8$). We further elaborate this by trying out model F_2, where interaction of age and income would be added. The results are shown in Table 14, but the Omnibus test indicates, that this interaction should not be added ($p = 0.064, \chi^2 = 3.423, df = 1$). Thus, we use the model F_1 as the final model.

The final model is thus is

$$y = 0.041x_1 - 0.655x_2 + 1.273x_3 + 0.071x_4 - 0.656$$

, where x_1 is age, x_2 and x_3 indicate low or high level of education and x_4 indicates the income in 15 categories.

The model is still poor, estimating only 8.3 % correctly of those who did not vote. The over all correctness is 85.2 %. However, this model only estimates 21 non voters, and making mistake in 5 cases. Thus, the correctness of non-votes is only 8.9 %. The Hosmer-Lemeshow test indicates that the model is good ($\chi^2 = 5.187, df = 8, p = 0.737$, which is well above 0.005), but we observed curios effect of the age in the models F_1 and F_2. When the interaction term of age and income was presented and controlled for, the

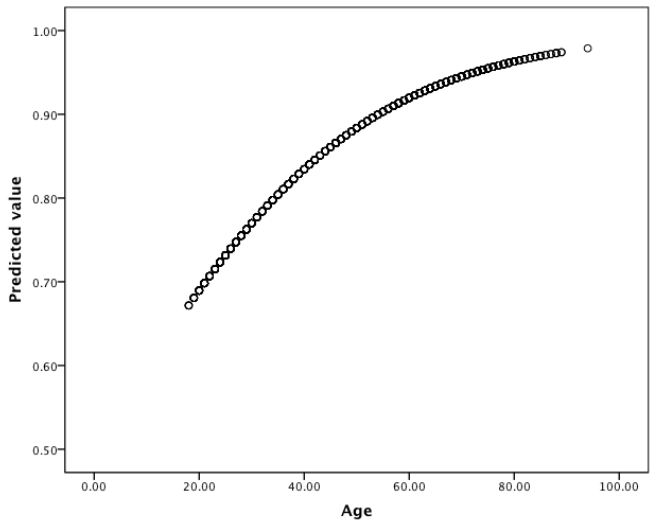

Figure 3: The effect of age to voting

significance of income itself became insignificant. Thus, there seems to be a connection between the age and the income even while the Omnibus test is just over the 5 %–cut-point (0.064). Let's illustrate the effect of age, income and the interaction of age and income.

In Figures 3 and 4 we see, that the age has a much much stronger effect to the predicted behaviour compared to the income. As we already discussed, the model F_2 showed, that there was a strong interaction of age and income and caused the income variable to be insignificant, even while the test measurements do not support adding that variable to the equation. Thus, even while the model F_1 indicates that the income is statistically significant, and thus supports the H_3, there is a reasonable doubt that the income variable is affected by the age of the responder. Secondly, as we see from the Figure 4, the effect of the income is not as strong as the effect of age. Thus, in Figure 5, we see the effect of both income and age in the predicted behaviour. Form that figure, we observe, that the effect of income is strongest in the young people and slowly becomes less significant. Thus, this may lead us to consider there may exists a collinearity problem, i.e. changes in age effect the changes in income. This may be observed in the income-variable of M_{F_2}, which is insignificant. Menard (2002, 75–78) discusses this problem and notes that "*[b]riefly, though, there is no really satisfactory solution to high collinearity.*" However, there is no significant correlation between the age and the income variables.

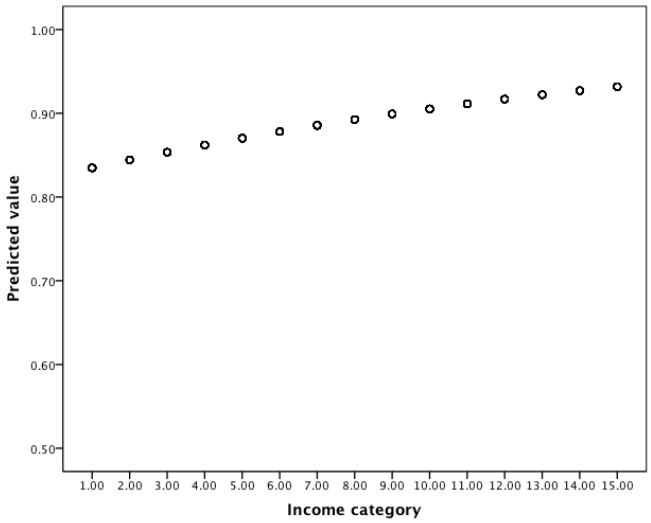

Figure 4: The effect of income to voting

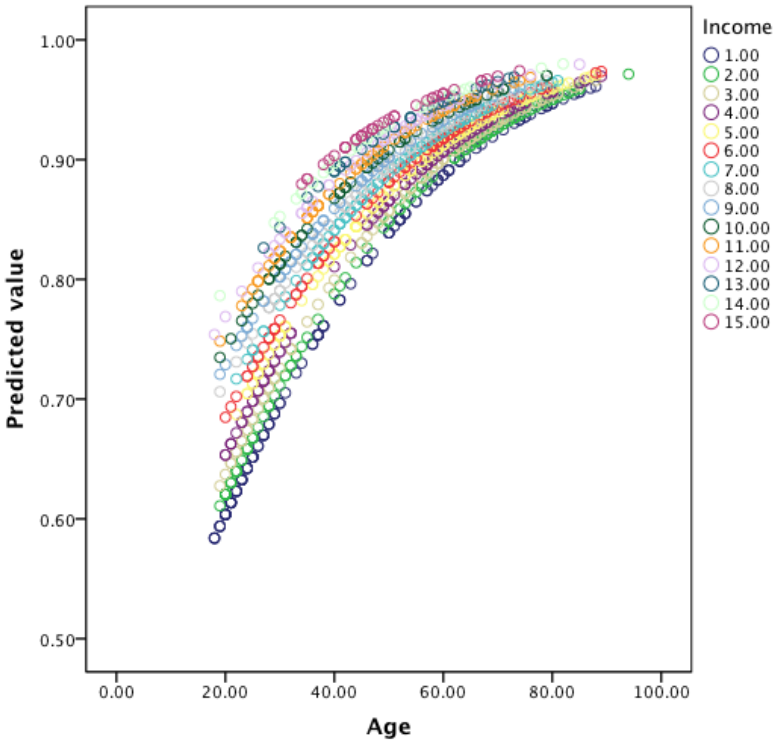

Figure 5: The effect of income and age to voting

Based on this examination, we clearly can accept both the H_1 and H_2. Especially the role of higher education is strong, as the odds-ratio is over three, meaning that those having higher education are two times more likely to vote compared to the middle and even more compared the lower levels of education. Also, as we have discussed above, the age is important. Some researchers, such as Wass (2007) and Topf (1995) discuss, that this is a generation effect in the voting, meaning that the younger generations are less active throughout their live in the election participation.

However, it is not clear if we should accept the hypothesis H_3. The significance of it in model F_1 shows strong support for this hypothesis, but as discussed above, there seems to be an interaction. The M_{F_2} indicates, that we should reject hypothesis H_3 and accept H_4 and the effect plot shown in Figure 5 indicates this also. Thus, in this work, even after the evidence supports accepting of H_3, the author does not accept it, but rather indicate that further research in the topic of income should be done, maybe reflecting also the social class, which has been used in previous research.

Lastly, overall the final model F_1 only predicts 8.3 % of the non-voters correctly. I.e. out of 192 non-votes only 16 are predicted correctly as non-voters. This suggest, that the background questions used here, sex, age, income and education are not enough to allow us make good predictions. This leads to the question if we can predict with background information only, or should the further research move to the use opinions, values and views more to allow building stronger models.

6 Conclusion

In this analysis we examined the effect of sex, age, education and income to voting behaviour. It may seem like this analysis is not well defined, having only tested four explanatory variables. However, it is in pair with the Finnish research done in this area. We used data an empiricalsample of 1218 voting aged Finns[8] , thus making the population all voting aged Finns. The data was collected by Taloustutkimus Oy and is stored in the Finnish Social Science Data Archive with the name '*Finnish National Election Study 2007*' (FSD 2269) [9].

[8]Missing 204 responders from the whole sample of 1422, mostly due to missing data in the question related to income.

[9]The author is responsible the analysis and results presented here.

We examined seven different models, including variables age (both linear and curve-linear), sex, income and education and interaction of age to income and sex. From these model, the model of age, income and education was selected as the final model, having total of 85.2 % correctness and good Hosmer & Lemeshow value ($p = 0.737, \chi^2 = 5.187, df = 8$). This model is the following equation

$$y = 0.041x_1 - 0.655x_2 + 1.273x_3 + 0.071x_4 - 0.656$$

, where x_1 is age, x_2 and x_3 indicate low or high level of education and x_4 indicates the income in 15 categories.

Based on previous research (Wass 2007, Martikainen et al. 2005), four hypothesis were selected under investigation. The hypothesis are

1. Older citizens vote more likely.
2. Those with higher education vote more likely.
3. Citizens with high income vote more likely.
4. An interaction exists between age and income.

Of these hypothesis, we found support for the first two. The last two however have mixed evidence, and thus more research would be required. More over, the over all model prediction power is not good, it can not explain in good level why some people vote and others do not. Thus, the other question needing examination is, if the use of only background variables is useful or should scholars of political science examine more values and believes as predictors.

References

Bernsteind, R., Chadha, A. & Montjoy, R. (2001), 'Overreporting voting: Why it happens and why it matters', *Public Opinion Quarterly* **65**(1), 22–44.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied logistic regression*, 2nd edn, John Wiley & Sons, Hoboken, New Jersey, USA.

Lewis-Beck, M. S. (1980), *Applied Regression. An introduction*, number 22 *in* 'Quantitative Applications in the Social Sciences', Sage Publications, London, United Kingdom.

Lijphart, A. (1997), 'Unequal participation: Democracy's unresolved dilemma', *The American Political Science Review* **91**(1), 1–14.

Martikainen, P., Martikainen, T. & Wass, H. (2005), 'The effect of socioeconomic factors on voter turnout in finland: A register-based study of 2.9 million voters.', *European Journal of Political Research* **44**(5), 645 – 669.

Martikainen, T. & Wass, H. (2007), *Äänestysaktiivisuuden ja puolueiden kannatuksen muutos vuoden 007 eduskuntavaaleissa Helsingissä*, number 8 *in* 'Tutkimuksia', Helsingin kaupungin tietokeskus.

Menard, S. (2002), *Applied Logistic Regression Analysis*, number 07-106 *in* 'Quantitative Applications in the Social Sciences', 2nd edn, Sage Publications, Thousand Oaks, California, USA.

Pampel, F. C. (2000), *Logistic regression. A primer*, number 132 *in* 'Quantitative Applications in the Social Sciences', Sage Publications, London, United Kingdom.

Silver, B. D., Anderson, B. A. & Abramson, P. R. (1986), 'Who overreports voting?', *American Political Science Review* **80**(2), xxxx.

Topf, R. (1995), Electoral participation, *in* H.-D. Klingemann & D. Fuchs, eds, 'Citizens and the state', Oxford University Press, Oxford, pp. 27–51.

van Deth, J. W. (2001), Studying political participation: Towards a theory of everything?, *in* 'Joint Sessions of Workshops of the European Consortium for Political Research'.

Wass, H. (2007), 'The effects of age, generation and period on turnout in finland 1975–2003', *Electoral Studies* **26**(3), 648–659.

Wattenberg, M. P. (2000), The decline of party mobilization, *in* R. J. Dalton & M. P. Wattenberg, eds, 'Parties without Partisans. Political Change in Advance Industrial Democracies', Oxford University Press, pp. 64–76.