EXAMINATION QUESTION SVSOS3003 Fall 2004 Some suggestions for answering the questions

Erling Berge
Department of sociology and political science,
Norwegian University of Science and Technology

"Read me"

In trying to answer the questions at the examinations it should be kept in mind that the questions often are problematic in relation to the model requirement of being based on the best available theory. The lack of theoretical foundation can be defended on two accounts. Most important is simply a lack of time and suitable data for construction of the "realistic" examination questions we want. But if it is taken for granted that the questions seldom are well grounded in theory, this will of course furnish the students will good arguments in the effort to critically evaluate the specification requirement of the models.

As you read the suggestions for answers presented here it is important to understand that it is not the only way of answering the questions. Most questions can be answered in many ways. Even if the technical questions have precise answers, the many evaluations necessary (e.g. "Is the distribution of the residuals close enough to the normal distribution for the tests to be believable?") are precisely evaluations. And for the evaluation, the arguments for or against are the essential parts of the answer.

During the examinations time is scarce. Few are able to answer exhaustively on all questions. In this note on the answers to the questions there has been done much more than we expect to find at the examination. Some sections contain more details of computation or stuff that may be relevant or related to the question asked, but not necessary to answer the question. However, the level of detail and additions varies.

I have to warn against presentation of errors and rash conclusions. This author has as much capacity to err as other people. Critical reading by students and colleagues is the best quality control there is. Anyone finding an error or thinking some other evaluation more appropriate is encouraged to write me, for example by e-mail: Erling.Berge@svt.ntnu.no, I will appreciate that.

QUESTION 1 (OLS-regression, weight 0,5)
In a Norwegian study of trust in fellow citizens differences between regions were investigated by OLS regression. Six models were estimated.

a) Explain what Model 1 tells about regional differences in trust.
b) Determine which of the six models best predicts the level of trust a person expresses. Find the F-statistic of a test of the best model against model 2.
c) Evaluate the hypothesis “The relationship between age and trust in fellow citizens is linear”. Find a 90% confidence interval for the impact of education in the best model.
d) Formulate the model identified as the best.
e) Based on the best model write up the formula for producing conditional effect plots according to age, sex and location in Oslo/Akershus or Trøndelag.
f) Discuss the degree to which the assumptions of OLS regression are met in the best model.

a) Explain what Model 1 tells about regional differences in trust.

The dependent variable is score on the trust index defined in the appendix tables. In model 1 we find the dummies of a regional code. We see that Oslo/ Akershus is the reference category. The regression coefficients are estimates of the expected difference between the reference category and the persons located in the indicated region. When all region variables have the value zero, the constant gives us the predicted value of the trust index for Oslo/ Akershus. On the scale from 0 to 10, the people of Oslo/ Akershus express an average trust of 6.435.

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	6.435	.075	85.914	.000		
	HedeOppl	.340	.135	2.512	.012	.759	1.317
	SouthEast	.028	.108	.257	.797	.636	1.572
	AgderRog	.030	.114	.267	.790	.667	1.499
	WestNorw	.192	.107	1.803	.072	.625	1.599
	Troendel	.334	.130	2.562	.010	.742	1.348
	NorthNor	.001	.128	.005	.996	.733	1.364

People living in Hedmark/ Oppland will on average score 0.34 points above those living in Oslo/ Akershus. The differences between those living in Oslo/ Akershus and those living in the South East of the country or in Agder/ Rogaland are negligible. The same is the difference between Oslo/ Akershus and North Norway. Those living in Trøndelag have a score 0.33 above those from Oslo/ Akershus or about the same as those living in Hedmark/ Oppland. For those living in West

Norway the model estimates a difference of 0.19 to those from Oslo/ Akershus. This is about half the difference of Trøndelag and Hedmark/ Oppland. However, it is not quite significant with a p-value of 0.072.

The picture emerging from this is that people living in Trøndelag/ Hedmark/ Oppland show a higher level of trust than the rest of the country except those living in West Norway where the level of trust probably lies somewhere in between.

b) Determine which of the six models best predicts the level of trust a person expresses. Find the F-statistic of a test of the best model against model 2.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.086(a)	.007	.004	1.482843	.007	2.494	6	1995	.021
2	.210(b)	.044	.040	1.456509	.037	19.198	4	1991	.000
3	.222(c)	.049	.044	1.452921	.005	10.847	1	1990	.001
4	.229(d)	.053	.046	1.451286	.003	3.243	2	1988	.039
5	.230(e)	.053	.046	1.451557	.000	.257	1	1987	.612
6	.232(f)	.054	.046	1.451518	.001	1.053	2	1985	.349

From the change statistics in the table “Model Summary” we see that the models 1, 2, 3, and 4 all are improvements on the previous model. Model 5 is not an improvement and probably neither is model 6 an improvements. Based on a criterion of parsimony model 4 will be seen as the best model.

But to be sure we may test model 6 against model 4.

$$F_{n-K}^{H} = \frac{\dfrac{RSS_{[K-H]} - RSS_{[K]}}{H}}{\dfrac{RSS_{[K]}}{n-K}} = ((4187{,}187 - 4182{,}206)/3)/(4182{,}206/(2002\text{-}17)) = 0{,}788$$

In the table of the F-distribution we see that with a true null hypothesis finding a $F^{3}_{1985} > 2.60$ will have a probability of less than 5%. The F-value of 0.788 found here will have a much larger probability than 0.05 and hence we reject the null hypothesis.

b) continued
Find the F-statistic of a test of the best model against model 2.

In a comparison of two models estimated on the same sample of n cases, one model with K parameters and one model with K-H parameters, the statistic

$$F_{n-K}^{H} = \frac{\dfrac{RSS_{[K-H]} - RSS_{[K]}}{H}}{\dfrac{RSS_{[K]}}{n-K}}$$

follows a F-distribution with H and n-K degrees of freedom if it is true that the H extra variables included in the big model have no effect (if H_0 “No impact of the new variables” is true) and the assumptions of OLS regression are met. In this formula the $RSS_{[K]}$ is the sum of squares of the residuals of the big model with K parameters (or K-1 variables) and $RSS_{[K-H]}$ is the sum of squared residuals in the small model where the H new variables are not included. We reject the null-hypothesis that the H new variables do not have an impact with level of significance α if F^{H}_{n-K} is larger than the critical value for level of significance α in the table of the F-distribution with H and n-K degrees of freedom.

ANOVA(g)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	32.906	6	5.484	2.494	.021(a)
	Residual	4386.651	1995	2.199		
	Total	4419.558	2001			
2	Regression	195.812	10	19.581	9.230	.000(b)
	Residual	4223.746	1991	2.121		
	Total	4419.558	2001			
3	Regression	218.709	11	19.883	9.419	.000(c)
	Residual	4200.849	1990	2.111		
	Total	4419.558	2001			
4	Regression	232.371	13	17.875	8.487	.000(d)
	Residual	4187.187	1988	2.106		
	Total	4419.558	2001			
5	Regression	232.913	14	16.637	7.896	.000(e)
	Residual	4186.645	1987	2.107		
	Total	4419.558	2001			
6	Regression	237.352	16	14.835	7.041	.000(f)
	Residual	4182.206	1985	2.107		
	Total	4419.558	2001			

Based on the ANOVA table we find that in comparing model 4 to model 2

$$F_{n-K}^{H} = \frac{\frac{RSS_{[K-H]} - RSS_{[K]}}{H}}{\frac{RSS_{[K]}}{n-K}} = ((4223.746 - 4187.187)/3) / (4187.187/(2002 – 14))$$

= (36,559/3) / (4187.187/1988) = 12,186 / 2,106 = 5,786

In the table of the F-distribution we see that with a true null hypothesis finding a F^{3}_{1989} > 2.60 will have a probability of less than 5%. The F-value of 5.789 found here will lead to rejection of the null hypothesis.

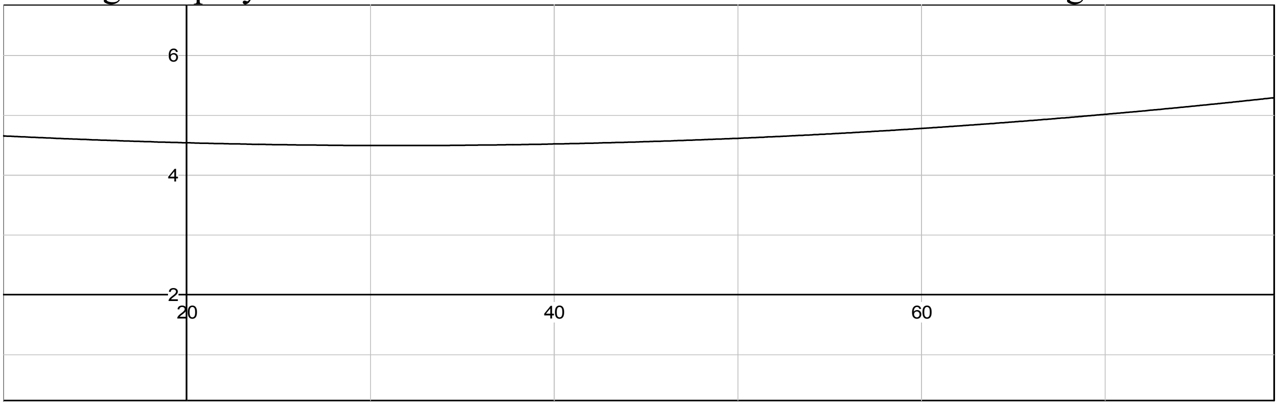
c) Evaluate the hypothesis "The relationship between age and trust in fellow citizens is linear". Find a 90% confidence interval for the impact of education in the best model.

The difference between model 2 and model 3 is the second order term in the age polynomial, Age2. The addition contributes significantly to the model with a t-value of 3.293 and a p-value of 0.001. This leads to a rejection of the null-hypothesis that the relationship between trust in fellow citizens and age is linear. The curvilinear relationship appears even stronger in model 4 where an interaction term between age and gender is introduced.

In model 3 age is related to the dependent variable through the relationship

$$4.842 - 0.022(\text{age}) + 0.00035(\text{age})^2$$

Plotting this polynomial we see that it reaches a low value around the age of 30



Applying calculus to the relationship

$$[d/d(\text{age})(4.842 - 0.022\text{age} + 0.00035\text{age}^2)] = 0$$

We find that for age = 31,43 trust reaches its lowest value.

c) continued
Find a 90% confidence interval for the impact of education in the best model.

In model 4 we see that b_{educ} = 0.107 with a standard error of 0.015. If we for the present computation assume that the residuals are normally distributed we can find a 90% confidence interval as

$b_{educ} – SE_{educ}*t_{0.1} < \beta_{educ} < b_{educ}+ SE_{educ}*t_{0.1}$

where b_{educ} is the estimated regression coefficient of education, SE_{educ} is the standard error, and $t_{0.1}$ is the critical value in the t-distribution in a two-sided test with α=0.1 and df = n-K = 2002 – 14 = 1988, degrees of freedom:

In table A4.1 in Hamilton (1992:350) we find that for df > 120 $t_{0.1}$ = 1.645. This means that a 90% confidence interval is given by

$0.107 – 0.015*1.645 < \beta_{educ} < 0.107 + 0.015*1.645$
$0.107 – 0.0247 < \beta_{educ} < 0.107 + 0.0247$
$0.0823 < \beta_{educ} < 0.1317$

In model 6 where education is included as a curvilinear variable, the computation of confidence interval is more complicated.

d) Formulate the model identified as the best.

To define a model there are three types of elements that need to be considered:

1. Definitions of the elements of the model (variables, error term, population and sample)
2. Definitions of the relationships among the elements of the model (the equation linking variables and error term, the sampling procedure linking sample and population, theories and time sequences of events and observations linking causes and effects)
3. Definitions of the assumptions that have to be met in order to use a particular method (such as the OLS method for linear regression) for estimating the model (model specification, distribution and properties of the error term)

At a minimum the formulation will include variable definitions, the formula linking variables and error term, and the assumptions needed to make valid inferences from the estimates of a particular procedure.
For the present problem we are told that data come from a random sample from the Norwegian population used by the European Social Survey in their 2002 investigation of the relations between institutional conditions and the attitudes,

values and opinions of citizens of European countries. Based on these data the following variables have been defined:

Y	**Trustindex**	
	Region in Norway	
X_1	HedmarkOppland	
X_2	SouthEast	
X_3	AgderRogaland	
X_4	WestNorway	
X_5	Troendelag	
X_6	NorthNorway	
	Age	
X_7	Age in years	Age
X_{11}	Age in years squared	Age2
X_8	**Female**	
	Education	
X_9	Education in years	Educ
X_{14}	Education in years squared	Educ2
X_{10}	**Number of household members**	NoHHmembers
	Interaction terms	
X_{12}	Female by age	FemaleAge
X_{13}	Female by age squared	FemaleAge2
X_{15}	Female by education	FemaleEduc
X_{16}	Female by education squared	FemaleEduc2

Since models 5 and 6 are not improvements on model 4, the variables X_{14} , X_{15} , and X_{16} are irrelevant and will not be considered further.

The main objective of the model is to investigate the regional variations in trust as measured by the Trustindex. If there in the Norwegian population is a linear or curvilinear relationship between the Trustindex and the defined independent variables we can write the equation linking the variables by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_8 X_{8i} + \beta_9 X_{9i} + \beta_{10} X_{10i} + \beta_{11} X_{11i} + \beta_{12} X_{12i} + \beta_{13} X_{13i} + \varepsilon_i \ .$$

Here ”i” runs over the whole of the Norwegian population. If we let k=0, 1, 2, 3, … ,13 β_k will be the unknown parameters showing how many measurement units

of y will be added to y per unit increase in X_k . ”ε_i” is the error term, a variable that comprises all relevant factors not observed as well as random noise in the measurement of y.
The equation for the model can also be written $y_i = E[y_i] + \varepsilon_i$
where $E[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} +...+ \beta_{13} x_{13i}$
(For $E[y_i]$ read “expected value of y_i”)

In Norway, 2036 persons were interviewed. A total of 34 persons have missing on one or more of the variables of the model and were removed. Since nothing more is indicated it must be assumed they were removed by listwise deletion. Only 2 were missing on the dependent variable, 32 were missing on education. This left 2002 for the analysis. There are no indications that the missing cases are non-random in relation to the dependent variable. Assuming that they are missing at random (MAR), listwise deletion is a proper procedure as long as it leaves enough cases to perform the analysis.

The Trustindex is a generalized measure of the confidence a persons puts in his or her fellow men. Attitudes as expressed in an interview situation are presumed to be based on more basic values shaped by socialization and accumulation of experiences within the social positions each individual have occupied. It is usually assumed that region will be a proxy for variations in such socialization and accumulated experiences. By the same reasoning it is also expected that age, gender and education will affect the general level of trust. It is also possible that the size of the household might affect this variable.

An OLS estimate of the model parameters defined above can be estimated as the b-values of $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} +...+ b_{13} x_{13i}$ that minimizes the sum of squared residuals,

$$RSS = \Sigma_i(y_i - \hat{y}_i)^2 = \Sigma_i e_i^2$$

(For ”$\hat{y}_i$” read “estimated” or “predicted” value of y_i or just “y-hat”.)

OLS estimates will be unbiased and efficient with a known sampling distribution if the following assumptions are true:

I: The model is correct, that is

- All relevant variables are included
- No irrelevant variables are included
- The model is linear in the parameters

II: The Gauss-Markov requirements for “Best Linear Unbiased Estimates” (BLUE)

- Fixed x-values (no random component in their measurement)

- The error terms have an expected value of 0 for all cases “i”
    - $E(\varepsilon_i) = 0$ for all “i”
- The error terms have constant variance for all cases “i” (homoscedasticity) for all “i”
    - $\text{var}(\varepsilon_i) = \sigma^2$ for all “i”
- The error terms do not correlate with each other across cases (no autocorrelation) for all “i” ≠ “j”
    - $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for all “i” ≠ “j”

III: The error terms are normally distributed

- The error terms are normally distributed (and with the same variance) for all cases for all “i”
    - $\varepsilon_i \sim N(0, \sigma^2)$ for all “i”

e) Based on the best model write up the formula for producing conditional effect plots according to age, sex and location in Oslo/Akershus or Trøndelag.

Model 4 has been identified as the best model.

B		Variable values	Conditions	Minimum	Maximum	Mean
5.334	(Constant)			,00	10,00	6,5425
.463	HedmarkOppland		0	,00	1,00	,0864
.120	SouthEast		0	,00	1,00	,1793
.134	AgderRogaland		0	,00	1,00	,1494
.276	WestNorway		0	,00	1,00	,1913
.434	Troendelag	1 or 0		,00	1,00	,0964
.103	NorthNorway		0	,00	1,00	,1014
-.046	Age	X		17,00	93,00	45,9361
-.814	Female	1 or 0		,00	1,00	,4580
.107	Educ		13	5,00	20,00	12,8761
.076	NoHHmembers		3	1,00	9,00	2,6723
.00059	Age2	X^2		289,00	8649,00	2401,2947
.051	FemaleAge	(1 or 0)X		25,00	400,00	170,9940
-.00051	FemaleAge2	(1 or 0)X^2		,00	93,00	21,1109

There is not said anything about Educ and NoHHmembers. To produce a conditional effect plot these variables are given reasonable values close to their average in the population. Educ is set to 13 and NoHHmembers to 3. Choosing other numbers will shift the regression line up or down a bit but not alter the pattern according to age.

One may express the conditional effect of both sex and location in one equation since Trøndelag is a dummy with Oslo/ Akershus as reference category.

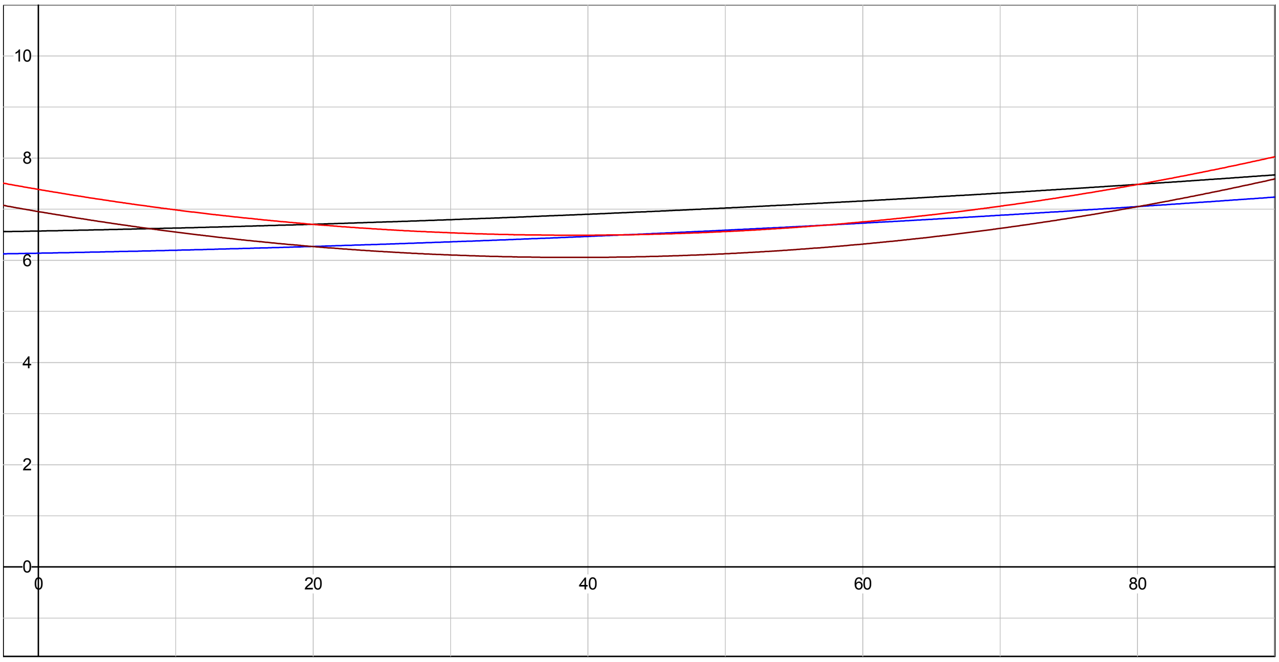
The conditional effect plot is produced by making a graph of
$Y = 5.334 + 0.434\text{Troendelag} – 0.046X – 0.814\text{Female} + 0.107*13 + 0.076*3 + 0.00059X^2 + 0.051\text{Female}*X – 0.00051\text{Female}*X^2 =$
$6.953+0.434\text{Troendelag} – 0.046X + 0.00059X^2 – 0.814\text{Female}+ 0.051\text{Female}*X – 0.00051\text{Female}*X^2$

There are however other ways of representing the relationships that may be more informative. The above expression may for example be restated as

$$Y=0.434\text{Troendelag}+(6,953–0.046X+0.00059X^2)+\text{Female}(– 0.814 + 0.051X – 0.00051X^2)$$

If Troendelag = 1 we get the graphs for Trøndelag and if Troendelag = 0 we get the graphs of Oslo/ Akershus (since Oslo/ Akershus is the excluded reference category). The trust index will for persons in Trøndelag be 0.434 points higher than for persons in Oslo/ Akershus.

$y=0.434\times1+(6.953-0.046x+0.00059x^2)+1\times(-0.814+0.051x-0.00051x^2)$
$y=0.434\times0+(6.953-0.046x+0.00059x^2)+1\times(-0.814+0.051x-0.00051x^2)$
$y=0.434\times1+(6.953-0.046x+0.00059x^2)+0\times(-0.814+0.051x-0.00051x^2)$
$y=0.434\times0+(6.953-0.046x+0.00059x^2)+0\times(-0.814+0.051x-0.00051x^2)$



From the diagram it appears that the difference between men and women in the trust they put in their fellow citizens are at its largest about the age of 50. The age effect among women is very close to linear while it is curvilinear for men.

Looking back at the formula it is seen that the FemaleAge2 coefficient nearly cancels the Age2 coefficient for men giving age a linear effect for women.

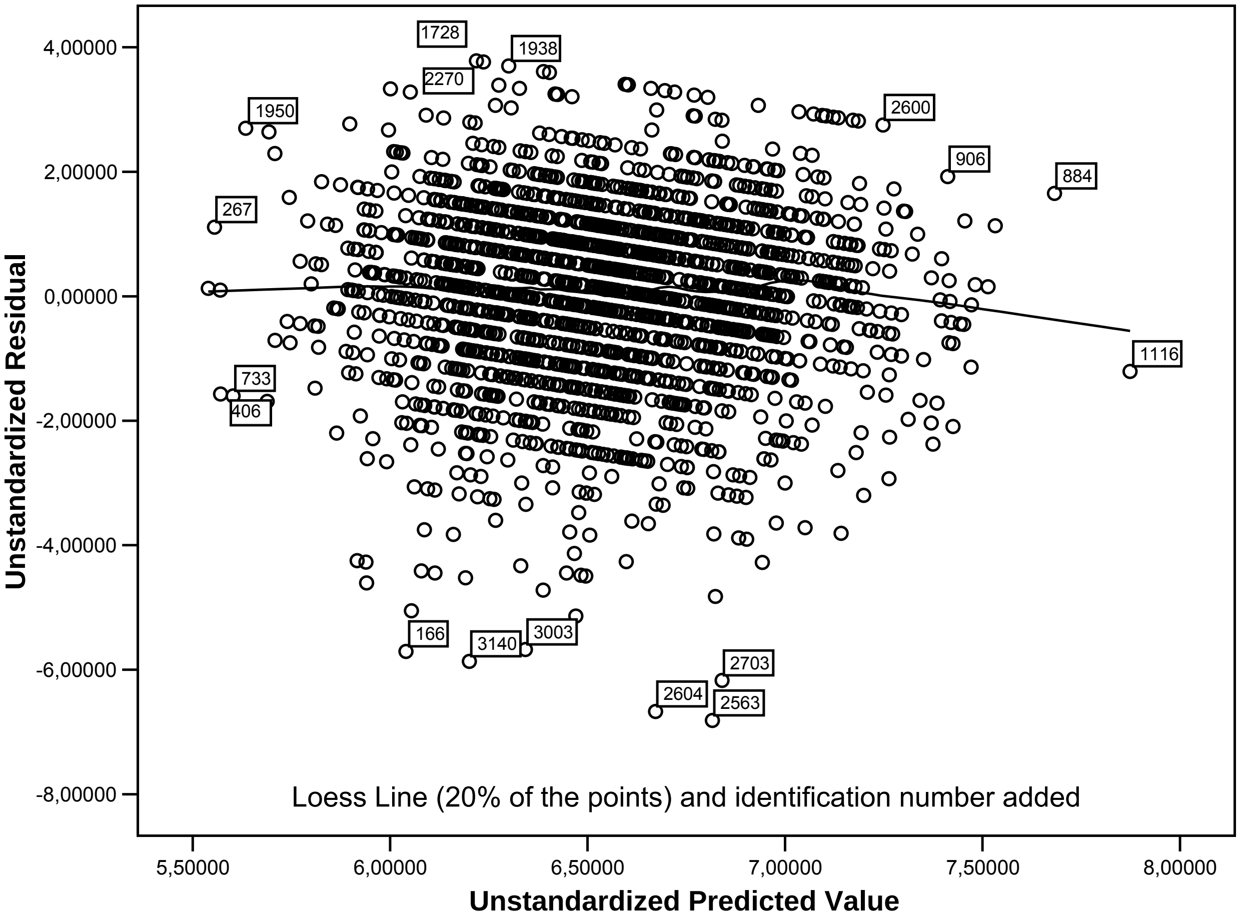
f) Discuss the degree to which the assumptions of OLS regression are met in the best model.

The assumptions that have to be met have been stated above. The first of the requirements, the specification of the model, cannot be investigated beyond stating that there are no irrelevant variables in the model and that it is linear in its parameters.

The Gauss-Markov requirements of fixed x-values and an expected value of 0 for the error terms cannot be tested. For the requirements of homoscedasticity and no autocorrelation there are tests.

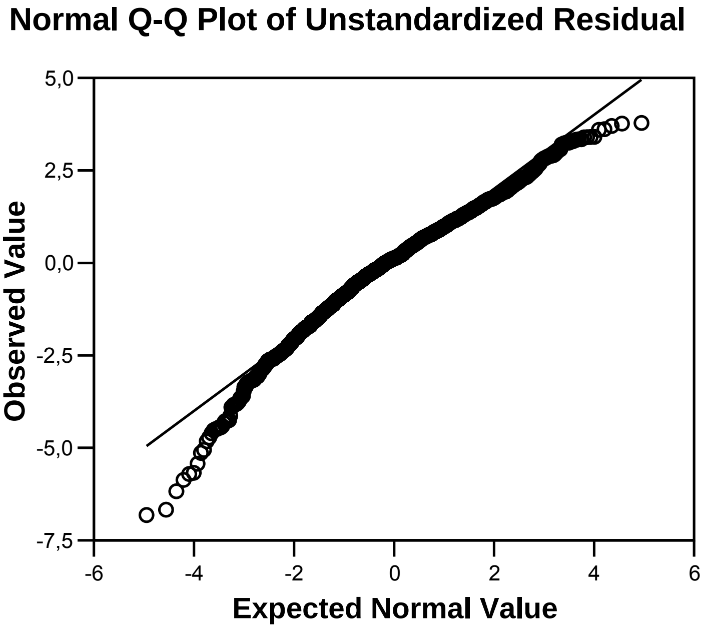
The sample is assumed to be a simple random sample of the Norwegian population. In such samples autocorrelation will not occur.

Homoscedasticity can be evaluated in a scatter plot of residual against predicted value:



The residuals are fairly evenly distributed around the mean value but with more negative than positive values. The distribution of the residual is slightly skewed. The loess line added to the plot dips down a bit at the upper values of predicted y possibly indicating a low level of heteroscedasticity. Both of these problems may be related to outliers such as the cases 2604, 2563, and 2703.

The assumption of normally distributed errors requires a symmetrical distribution of the residuals. From the scatter plot above we see that is not the case. The distribution can further be investigated in a comparison of the quantiles of the distributions of the residuals to the quantiles of the normal distribution:

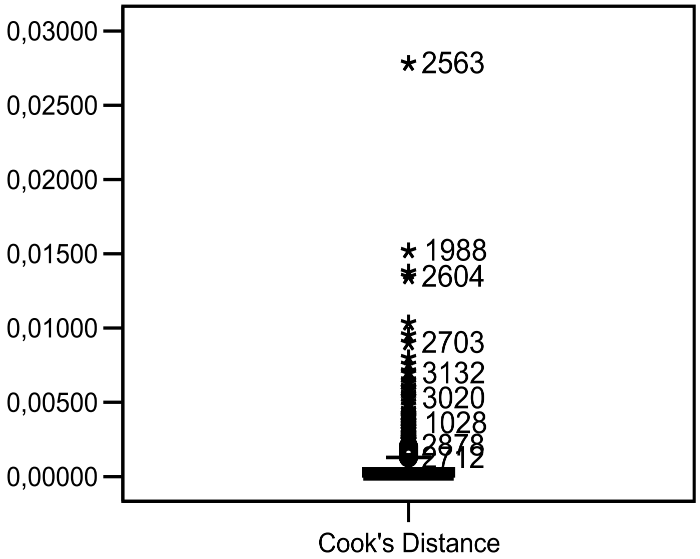


The plot confirms our conclusions above. The distribution of the residuals is negatively skewed with a few negative outliers. This casts doubt on the tests performed. They are not trustworthy. Maybe model 4 is not the best model after all.

One of the problems may be the outliers. The three largest residuals are found for respondents no

Respondent's identification number	TrustIndex	Predicted Value	Residual
2563	,00	6,7964	-6,79642
2604	,00	6,6035	-6,60352
2703	,67	6,8194	-6,15273

Looking at the DFBETAs it is seems that respondents with id no 2703, 2604, 1988 are candidates for possibly unwanted high impact. Cook’s D indicates respondent no 2563 as the one with highest influence.

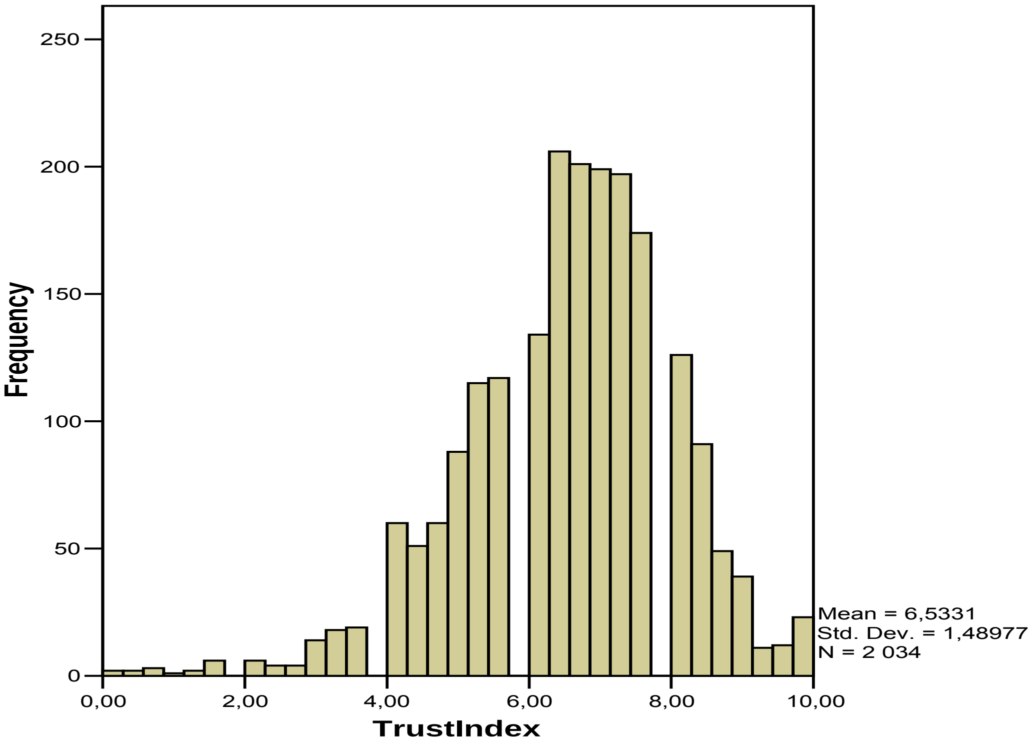


Looking closer at the most influential case, 2563, we see it is a 84 year old lady living in West Norway without any trust in her fellow citizens at all.

idno	2563	2604	2703
TrustIndex	0,000	0,000	0,667
HedmarkOppland	0	0	0
SouthEast	0	0	0
AgderRogaland	0	0	0
WestNorway	1	0	0
Troendelag	0	1	1
NorthNorway	0	0	0
Age	84	66	44
Female	1	1	1
Educ	9	9	12
NoHHmembers	1	1	3
RES_1	-6,816	-6,672	-6,174
ZRE_1	-4,697	-4,597	-4,254
COO_1	0,028	0,013	0,009
LEV_1	0,017	0,008	0,006
DFB4_1 (WestNorway)	-0,018	0,001	0,001
DFB5_1 (Troendelag)	0,000	-0,034	-0,032
DFB8_1 (Female)	-0,106	0,014	0,025

Many of us have encountered such persons. There do not seem to be any invalid data, and hence no reason to exclude cases. The most obvious thing the three persons listed here have in common is a very low trust in their fellow citizens.

From the distribution of the trust index it is seen that there are very few persons with a value below 3.



With non-normal residuals and a distribution like this a transformation of the dependent variable may be a solution to the problem of valid tests. For example the square root of the trustindex value or even the natural logarithm of it may be candidates.

QUESTION 2

In the same study of trust, regional differences in viewing politicians as vote maximizers were also investigated. Based on the information available in the tables attached, please answer the following:

a) Discuss what the study says about the hypotheses
H1: The relationship between “Politicians interested in votes rather than peoples opinions” and “Age” is curvilinear
H2: The impact of “Age” on “Politicians interested in votes rather than peoples opinions” depends on the sex of the person

b) Based on model 4 write up the equation determining the relationship between dependent and independent variables in the population studied and present the assumptions that have to be met to draw valid inferences from the estimated relationships

c) Find in model 2 the odds ratio between women and men for thinking that politicians are vote maximizers. Discuss regional variation in “Politicians interested in votes rather than peoples opinions”

d) Discuss the degree to which the assumptions of logistic regression have been met in model 5 estimated for this question

e) Based on model 4 write up the equation for a conditional effect plot according to age of respondent that will maximise the probability of observing Y=1 on the variable “Politicians interested in votes rather than peoples opinions”

f) Find from model 5 an expression for the odds ratio for observing Y=1 on “Politicians interested in votes rather than peoples opinions” between groups of men with one year difference in age

a) Discuss what the study says about the hypotheses
H1: The relationship between “Politicians interested in votes rather than peoples opinions” and “Age” is curvilinear
H2: The impact of “Age” on “Politicians interested in votes rather than peoples opinions” depends on the sex of the person

It is assumed that the hypotheses apply to the logit relationships.

H1 is then tested by comparing the Model of block 5 to the model estimated in block 4. The test statistic is

$$\chi^2_H = -2\{\log_e \mathcal{L}_{K-H} - \log_e \mathcal{L}_K\}$$

In this test H=2, K=16, $-2\log_e \mathcal{L}_K = 2302{,}313$, and $-2\log_e \mathcal{L}_{K-H} = 2306{,}050$

This means that $\chi^2_H = 2306{,}050 - 2302{,}313 = 3{,}737$. The critical value of the Chi-square distribution with 2 degrees of freedom and a level of significance 5% is 5.991. The two terms with age squared do not add significantly to the model.

Hence we have to reject the H1 hypothesis and conclude that age is linearly related to y.

The same result follows from the omnibus test of Block 5. The test statistic of block 5 in the omnibus test of model coefficients is 3.737. The p-value is given as 0.154 and with a significance level of the test of 0.05 we have to reject H1.

H2 is tested by including an interaction term between age and sex in the model. But since sex also is included in an interaction term with education, the level of multicollinearity will increase and t-tests of the significance of the interaction term alone will not be precise enough. The two variables Age and FemaleAge included in block 4 are significant together since $\chi^2_H = -2\{\log_e \mathcal{L}_{K-H} - \log_e \mathcal{L}_K\}$ = 2328,464 – 2306,050 = 22,414 is larger than the 5.991 critical level for 2 degrees of freedom (H=2). Hence we have to reject a null hypothesis of no impact of Age and FemaleAge. Again, the same result follows from the omnibus test of Block 4.

Until further investigations show otherwise we accept H2.

b) Based on model 4 write up the equation determining the relationship between dependent and independent variables in the population studied and present the assumptions that have to be met to draw valid inferences from the estimated relationships

In the population investigated there is assumed to be a logistic relationship between the probability of a value of 1 on the dependent variable Y and the independent X-variables. The model can be written as

$$\Pr[Y_i=1] = E[Y_i],$$

where $Y_i=[1/(1+\exp\{-L_i^*\})] + \varepsilon_i$, i=1, … ,n, ε_i is the error term and L_i is the logit defined as

$$L_i = \beta_0 + \sum_{k=1}^{K-1} \beta_k X_{ki}$$

where k is an index indicating the K-1 independent variables. The relationship between dependent and independent variables is supposed to be valid for all individuals “i” in the Norwegian population. For model 4 the following variables can be defined:

X_1	HedmarkOppland
X_2	SouthEast
X_3	AgderRogaland
X_4	WestNorway
X_5	Troendelag
X_6	NorthNorway
X_7	Voted
X_8	LeftRightScale
X_9	Female
X_{10}	Educ
X_{11}	FemaleEduc
X_{12}	Age
X_{13}	FemaleAge

With K=14 and "I" running over the sample of n=1954 persons, the parameters of this model can be estimated by the maximum likelihood method, and valid inferences can be made if the following assumptions are met:

- The model is correctly specified, i.e.:
  - All conditional probabilities for Y=1 are logistic functions of the x-variables (this means the logit is linear in its parameters)
  - There are no irrelevant variables included in the model
  - There are no relevant variables excluded from the model
- All independent variables have been measured without errors
- All cases are independent

In addition it should be observed that the method also require

- No perfect multicollinearity
- No perfect discrimination

And that the precision of the estimates are affected by

- High degree of multicollinearity
- High degree of discrimination
- Small sample

If the assumptions are met, the estimates of the parameters will be unbiased, efficient (minimum variance) and normally distributed. The likelihood ratio test can be used and in large samples b_k/ SE_{bk} will asymptotically follow a normal distribution.

c) Find in model 2 the odds ratio between women and men for thinking that politicians are vote maximizers. Discuss regional variation in “Politicians interested in votes rather than peoples opinions”

The odds ratio between women and men is defined as the ratio of exp(Logit(women))/exp(Logit(men)). In model two the odds ratio between women and men is found to be OR(women/men) = exp(-0,279) = 0,757. This means that the odds for finding Y=1 among women are 24,3% less than for comparable men.

If we try to determine the odds ratio between women and men in model 3 it becomes a bit more complicated. In that case it will be best to use the definition of the odds ratio.
In model 3 the logit is defined as
$L_i = 1{,}092 - 0{,}646\ \text{HedmarkOppland}_{i1} + 0{,}168\ \text{SouthEast}_{i2} - 0{,}084\ \text{AgderRogaland}_{i3} -{,}320\ \text{WestNorway}_{i4} - 0{,}408\ \text{Troendelag}_{i5} - 0{,}003\ \text{NorthNorway}_{i6} - 0{,}296\ \text{Voted}_{i7} + 0{,}026\ \text{LeftRightScale}_{i8} + 1{,}148\ \text{Female}_{i9} - 0{,}118\ \text{Educ}_{i10} - 0{,}118\ \text{FemaleEduc}_{i11}$
The we find the odds as
$\text{Odds(women)} = \exp\{\ 1{,}092 - 0{,}646\ \text{HedmarkOppland}_{i1} + 0{,}168\ \text{SouthEast}_{i2} - 0{,}084\ \text{AgderRogaland}_{i3} -{,}320\ \text{WestNorway}_{i4} - 0{,}408\ \text{Troendelag}_{i5} - 0{,}003\ \text{NorthNorway}_{i6} - 0{,}296\ \text{Voted}_{i7} + 0{,}026\ \text{LeftRightScale}_{i8} + 1{,}148\ \text{Female}_{i9} - 0{,}118\ \text{Educ}_{i10} - 0{,}118\ \text{FemaleEduc}_{i11}\ \}$
Hence we find the odds ratio of women to men as OR (women/men) =

$$\frac{e^{1{,}092 - 0{,}646\text{HedOpp} + 0{,}168\text{SEast} - 0{,}084\text{AgdRog} -{,}320\text{WNor} - 0{,}408\text{Troend} - 0{,}003\text{NNor} - 0{,}296\text{Vot} + 0{,}026\text{LRScale} + 1{,}148(\text{Female}=1) - 0{,}118\text{Educ} - 0{,}118(\text{Female}=1)\text{Educ}}}{e^{1{,}092 - 0{,}646\text{HedOpp} + 0{,}168\text{SEast} - 0{,}084\text{AgdRog} -{,}320\text{WNor} - 0{,}408\text{Troend} - 0{,}003\text{NNor} - 0{,}296\text{Vot} + 0{,}026\text{LRScale} + 1{,}148(\text{Female}=1) - 0{,}118\text{Educ} - 0{,}118(\text{Female}=1)\text{Educ}}}$$

With elements of the logit that are the same for women and men collected as Const the odds ratio becomes

$$OR$$

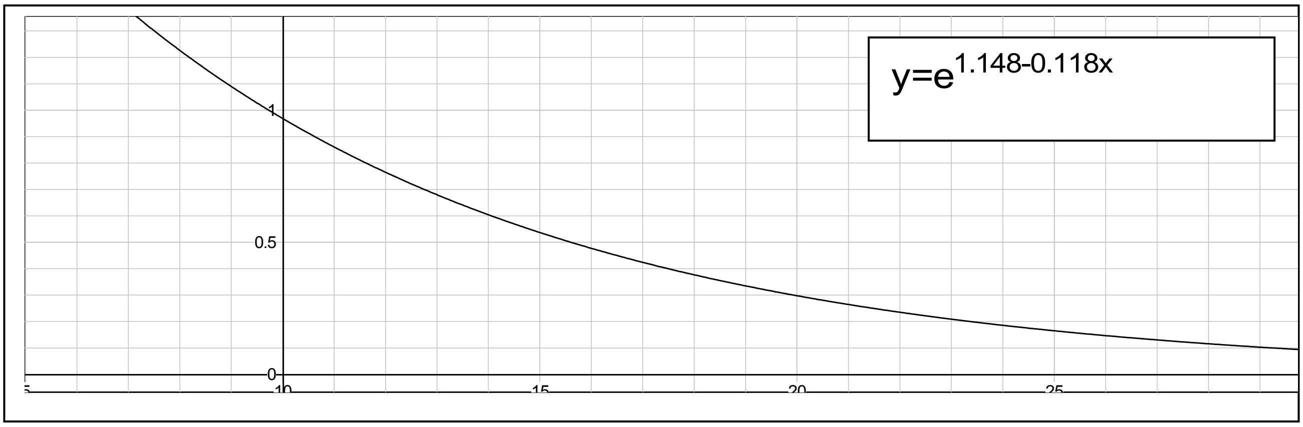
$$= \frac{e^{\text{Const} + 1{,}148(\text{Female}=1) - 0{,}118(\text{Female}=1)\text{Educ}}}{e^{\text{Const} + 1{,}148(\text{Female}=0) - 0{,}118(\text{Female}=0)\text{Educ}}}$$

$$= \frac{e^{1{,}148(\text{Female}=1) - 0{,}118(\text{Female}=1)\text{Educ}}}{e^{0}}$$

$$= e^{1{,}148(\text{Female}=1) - 0{,}118(\text{Female}=1)\text{Educ}}$$

$$= e^{1{,}148 - 0{,}118\text{Educ}}$$

The odds ratio between women and men in model 3 depends on the amount of education they have. From the plot below we see that women with more than 9 year of education increasingly have smaller odds for thinking politicians are vote maximizers compared to men with the same amount of education.



In model 4 we will get an additional term making the odds ratio dependent on both education and age.

c) continued:
2 c) continued
Discuss regional variation in “Politicians interested in votes rather than peoples opinions”
Since the first part of 2c is restricted to model 2 it may be a reasonable assumption that this part should be equally restricted. That is not a necessary implication but must be accepted. The comments about regional variations can be made both based on odds ratios and based on logit coefficients. The conclusions in broad terms will be the same. The comments here are based on logit coefficients for all models.

The estimates of the logit coefficients of the region dummies tell us how different they are on average, the persons living in this region compared to persons in the reference region, Oslo/ Akershus. Only Hedmark/ Oppland is at significance level 0.05 different from the reference region consistently across all models.

Variables in the Equation

	B block 1*	B block 2	B block 3	B block 4	B block 5
HedmarkOppland	**-,435**	**-,444**	**-,646**	**-,655**	**-,638**
SouthEast	**,335**	**,329**	,168	,176	,184
AgderRogaland	,098	,079	-,084	-,052	-,044
WestNorway	-,147	-,166	**-,320**	**-,282**	**-,294**
Troendelag	-,258	-,286	**-,408**	**-,391**	**-,385**
NorthNorway	,152	,141	-,003	,014	,021

* Note that **boldface** means significant at level 0.1 and **boldface** means significant at level 0.05

The sign of the coefficient tells if the probability of thinking politicians are vote maximizers increases (+) of decreases (-). In model 2 only SouthEast gives a higher probability than Oslo/ Akershus of observing Y=1 and the difference is significant at level 0.05. Likewise Hedmark/ Oppland gives a lower probability significant at level 0.05.

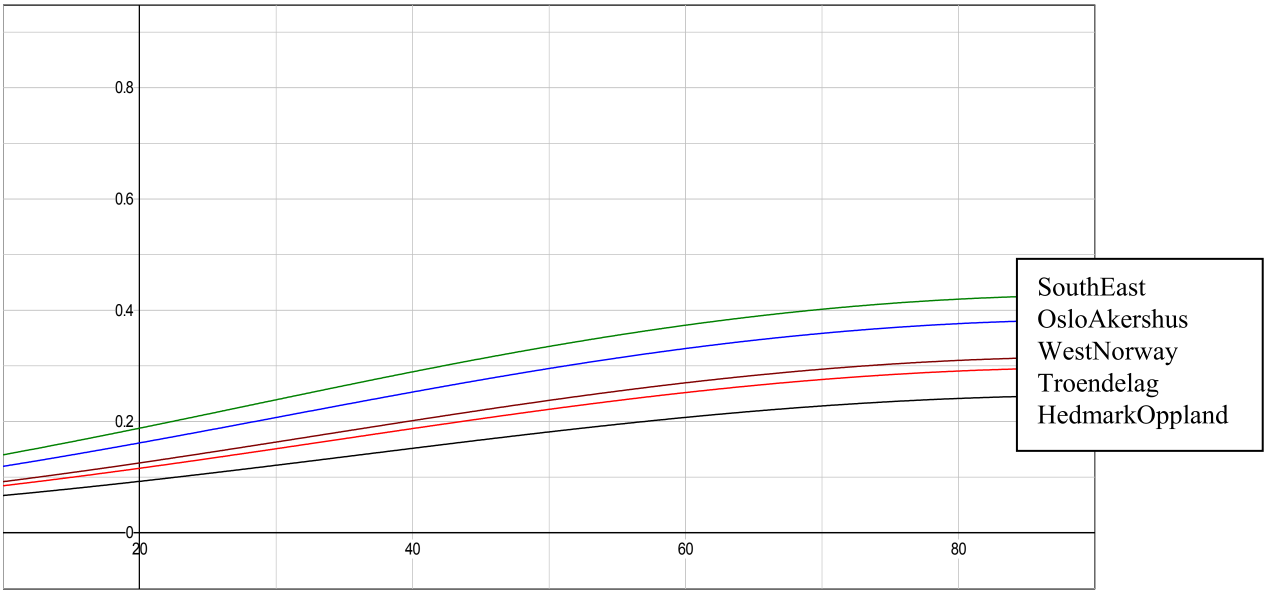
As variables are added we see that the impact of region changes. The addition of “voted”, LeftRightScale” and “Female” in block 2 increases the difference between Hedmark/ Oppland and the reference category Oslo/ Akershus by about 50%. And when education and the interaction between gender and education are added in block 3, the South East region ceases to be significantly different from Oslo/ Akershus while West Norway and Trøndelag becomes significantly different. From block 3 to block 5 the coefficient estimates do not change much.

It is interesting to note that the basic pattern of trust/ confidence in politicians is the same here as in the more general trust question of 1a.

We know that education and political sympathies are not evenly distributed across regions. Hence it must be expected that to assess the true impact of region, political sympathies and education has to be controlled for. We also know that gender and age is evenly distributed across regions and should not affect estimates of the impact of region. The pattern of changes of the coefficient estimates is consistent with this.

Based on the estimates from model 5 the pattern of the coefficients suggests two regions with somewhat different views on politicians. Persons living in Oslo/ Akershus, Sout East, Agder/ Rogaland, and North Norway share one view while those living in Hedmakr/ Oppland, Troendelag and West Norway share another view with a lower probability of thinking politicians are vote maximizers.

In the different regions the probabilities of thinking politicians are vote maximizers for women that voted, with 13 year of education and placing themselves at 6 on the left-right scale are shown below. The left out regions Agder/ Rogaland and North Norway are very close to Oslo/ Akershus.



$y=1/(1+e^{-(1.063-0.638\times1+0.184\times0-0.044\times0-0.294\times0-0.385\times0+0.021\times0-0.477\times1+0.032\times6-0.703\times1-0.09\times13-0.1\times1\times13-0.023\times x+0.065\times1\times x+0.00035\times x^2-0.00058\times1\times x^2)})$	SouthEast
$y=1/(1+e^{-(1.063-0.638\times0+0.184\times0-0.044\times0-0.294\times0-0.385\times0+0.021\times0-0.477\times1+0.032\times6-0.703\times1-0.09\times13-0.1\times1\times13-0.023\times x+0.065\times1\times x+0.00035\times x^2-0.00058\times1\times x^2)})$	OsloAkershus
$y=1/(1+e^{-(1.063-0.638\times0+0.184\times0-0.044\times0-0.294\times0-0.385\times1+0.021\times0-0.477\times1+0.032\times6-0.703\times1-0.09\times13-0.1\times1\times13-0.023\times x+0.065\times1\times x+0.00035\times x^2-0.00058\times1\times x^2)})$	WestNorway
$y=1/(1+e^{-(1.063-0.638\times0+0.184\times0-0.044\times0-0.294\times1-0.385\times0+0.021\times0-0.477\times1+0.032\times6-0.703\times1-0.09\times13-0.1\times1\times13-0.023\times x+0.065\times1\times x+0.00035\times x^2-0.00058\times1\times x^2)})$	Troendelag
$y=1/(1+e^{-(1.063-0.638\times0+0.184\times1-0.044\times0-0.294\times0-0.385\times0+0.021\times0-0.477\times1+0.032\times6-0.703\times1-0.09\times13-0.1\times1\times13-0.023\times x+0.065\times1\times x+0.00035\times x^2-0.00058\times1\times x^2)})$	HedmarkOppland

d) Discuss the degree to which the assumptions of logistic regression have been met in model 5

The assumptions have been stated above. The discussion will assume that the tables refer to model 5. It has already been determined that in model 5 Age2 and FemaleAge2 are irrelevant variables. From the estimate of model 5 it also is apparent that the LeftRightScale do not contribute significantly to the model and must be considered irrelevant based on the information presented here. However, there may conceivably be good theoretical arguments for keeping the LeftRightScale in the model. The possible absence of relevant variables cannot be commented upon.
The test of age as a curvilinear variable in the logit rejected the possibility of curvilinearity for age. The possibility that Educ and LeftRightScale might be curvilinearly related to the logit ought to be investigated.

Plots of the logits computed in the table of probability of ln(p/(1-p)) according to LeftRightScale, Age and Education do not give clear indications of curvilinearity. The most probable curvilinear relationship is for age, but this has already been rejected.

Chart of Logit for groups on the LeftRightScale

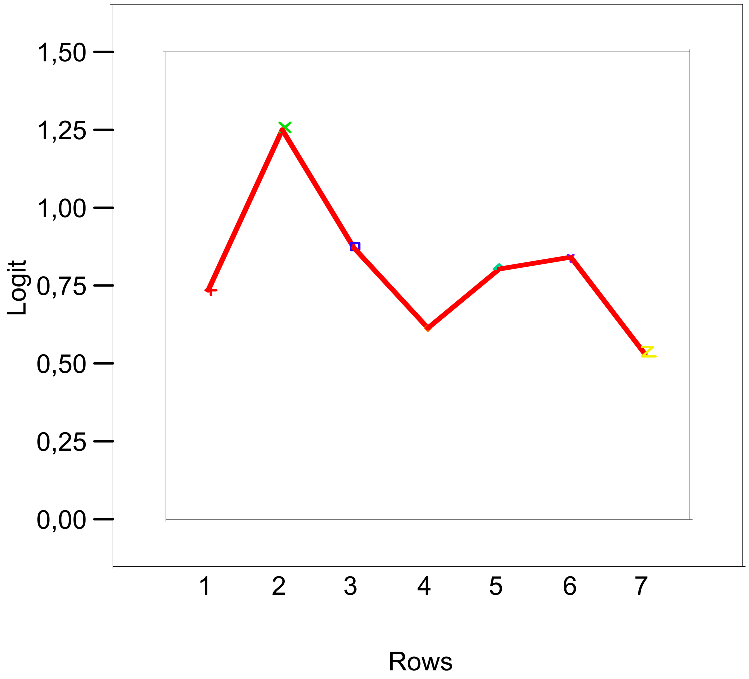


Chart of Logit for Age groups

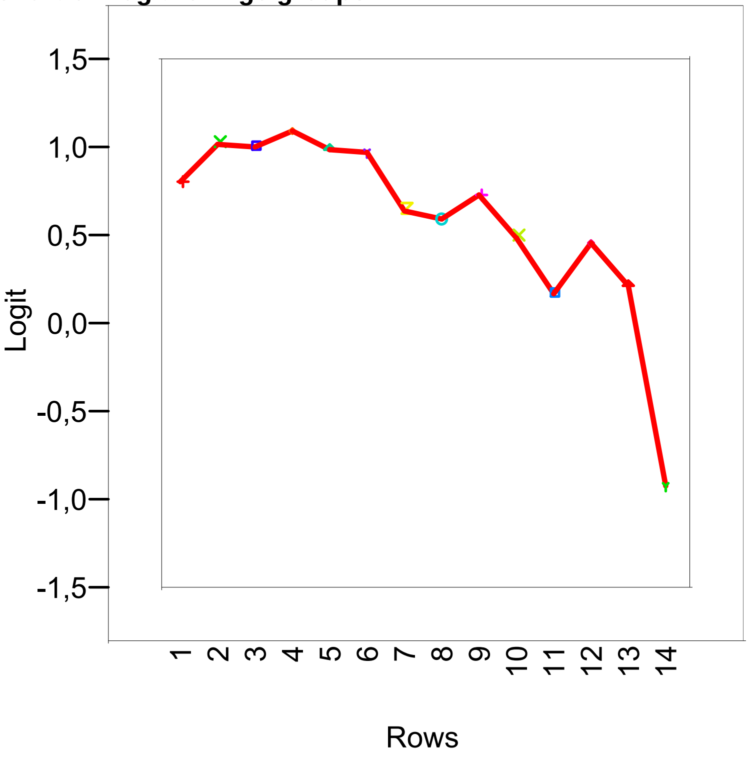
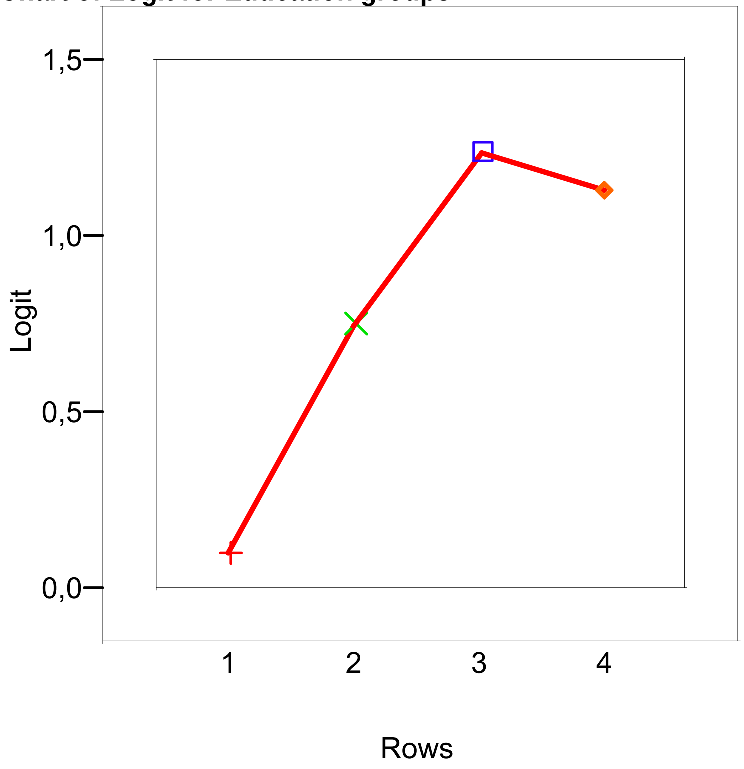


Chart of Logit for Education groups



The conclusion on the specification requirement must be that there are 3 irrelevant variables. In other respects the model must be seen as correct.

The two other requirement that cases are independent and that the x-variables are measured without error cannot be tested. But the source of the data, The European Social Survey assures that the data have been collected using the best available procedures.

The absence of perfect multicollinearity and perfect discrimination is demonstrated by the existence of the model estimates.

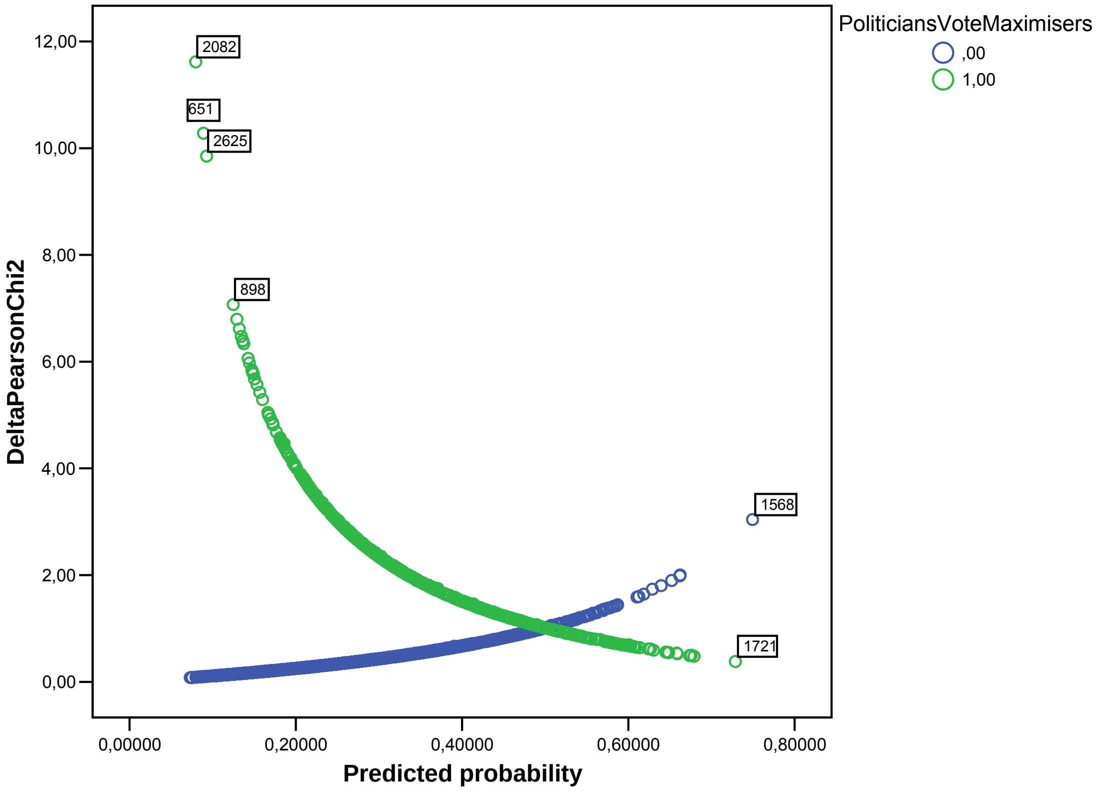
Strictly speaking the above discussion covers the assumptions of the model. But the possibility for severe statistical problems due to multicollinearity, discrimination, small sample and influential cases might also be investigated.

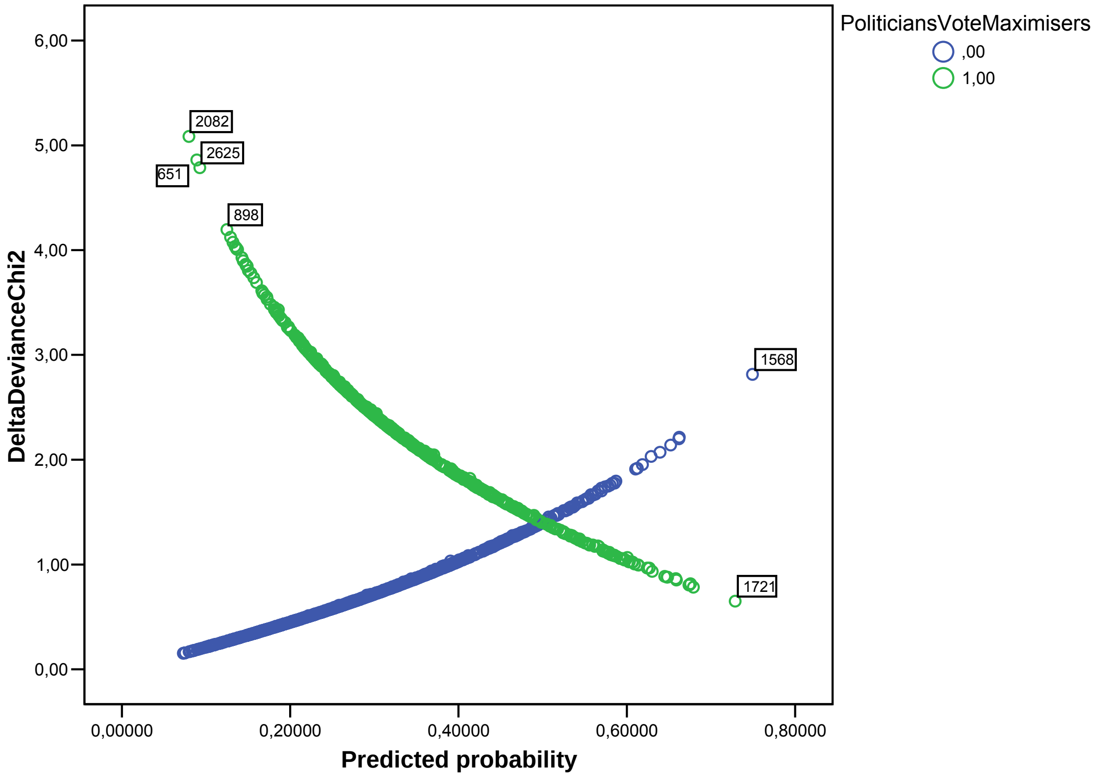
The effective sample is 1954 cases, and the dependent variable has about 31,5% in the Y=1 category. The sample size should be more than large enough.

There are no cross tabulations of the dependent variable and the dichotomous independent variables. Hence the possibility for some degree of discrimination cannot be commented upon.

The level of multicollinearity is high in model 5 due to interaction terms and age squared as a part of the test of age as a curvilinear variable. As long as this is considered in the test procedure employed, multicollinearity due to model specification cannot be considered a problem.

The possible existence of influential cases can be investigated in the plots of the DeltaPearsonChi2 and the DeltaDevianceChi2. As a rule of thumb values of these “poorness of fit” statistics above 4 are considered “significant” since their distribution is asymptotically χ^2 with 1 degree of freedom. The ultimate test of influence is to compare two regressions, one with the possibly influential case included and one with it excluded. Based on the figures below 3 cases may be considered for such an investigation: cases no 2082, 2625, and 651. From these cases and down to the next in line there is a clear “gap”. From inspection of the plot of the analogue of Cook’s D statistic a fourth case no 884 may be added.





The four cases identified above seem to have several characteristics in common. They are all female with high education. They have all voted and they all consider politicians to be vote maximizers.

idno	651	884	2082	2625
Age	31	82	35	23
Female	1	1	1	1
Educ	18	15	18	15
OsloAkershus	1			
HedmarkOppland		1		
SouthEast				
AgderRogaland				
WestNorway			1	
Troendelag				1
NorthNorway				
Politicians-VoteMaximisers	1	1	1	1
Voted	1	1	1	1
LeftRightScale	4	7	6	7

The 884 case has the highest analogue to Cook's D statistic; the other three are identified by high values on the DeltaPearsonChi2 and the DeltaDevianceChi2 statistics. Regression where these two groups of cases were deleted should have been compared to the reported regression to determine their actual impact.

e) Based on model 4 write up the equation for a conditional effect plot according to age of respondent that will maximise the probability of observing Y=1 on the variable “Politicians interested in votes rather than peoples opinions”

B		Minimum	Maximum	Value maximizing logit
,514	Constant			
-,655	HedmarkOppland	,00	1,00	0
,176	SouthEast	,00	1,00	1
-,052	AgderRogaland	,00	1,00	0
-,282	WestNorway	,00	1,00	0
-,391	Troendelag	,00	1,00	0
,014	NorthNorway	,00	1,00	0
-,494	Voted	,00	1,00	0
,032	LeftRightScale	,00	10,00	10
,254	Female	,00	1,00	1
-,102	Educ	5,00	20,00	5
-,080	FemaleEduc	,00	18,00	5
,011	Age	17,00	93,00	x
,009	FemaleAge	,00	93,00	x

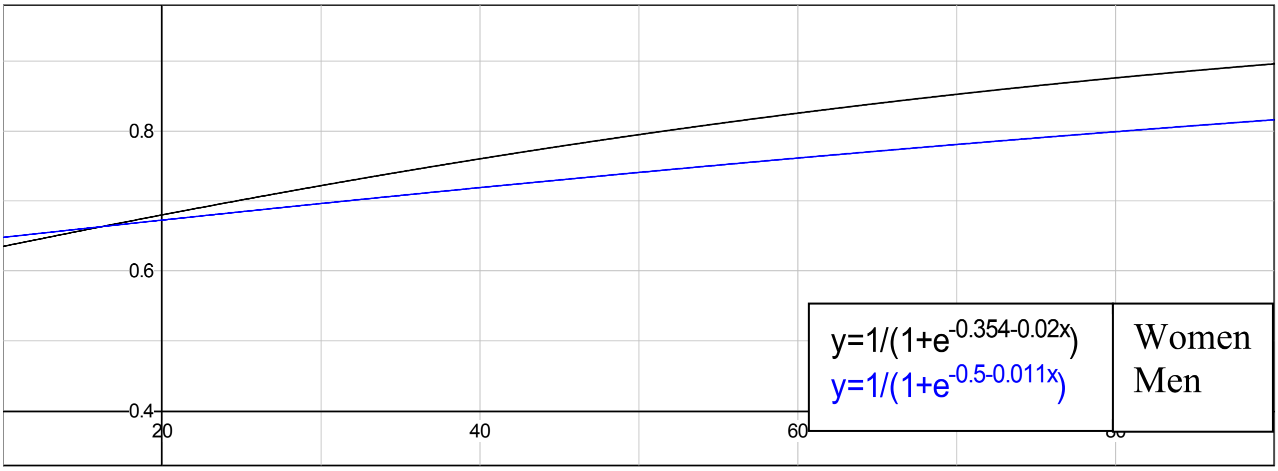
To maximize the probability one has to choose variable values maximizing the logit. That means choosing maximum variable values where the coefficient is positive and minimum values where the coefficient is negative. It is seen that the person maximizing will be a non-voter placing him or herself at the extreme right on the LeftRightScale and located in the Southeast region. Female, age and education are linked together and it is not at the outset obvious how to maximize the logit for these. Since the coefficient for both Educ and FemaleEduc is negative, Educ should have its minimum value of 5. And since the coefficients for Female, Age and FemaleAge are all positive one should choose the4 highest value of Female (female=1).

Hence a person maximizing the probability of thinking politicians are vote maximizers will be a female person with a minimum of education living in the SouthEast who do not vote but place herself at the extreme right of the LeftRightScale.

The equation for the logit maximizing the probability according to age can then be written

$L_i = 0{,}514 + 0{,}176(\text{SouthEast}_i) + 0{,}032(\text{LeftRightScale}_i)$
$+ 0{,}254(\text{Female}_i) - 0{,}102(\text{Educ}_i) - 0{,}08(\text{Female}_i * \text{Educ}_i)$
$+ 0{,}011(\text{Age}_i) + 0{,}009(\text{Female}_i * \text{Age}_i) =$
$0{,}514 + 0{,}176(\text{SouthEast}_i = 1) + 0{,}032(\text{LeftRightScale}_i = 10)$
$+ 0{,}254(\text{Female}_i = 1) - 0{,}102(\text{Educ}_i = 5) - 0{,}08(\text{Female}_i = 1)(\text{Educ}_i = 5)$
$+ 0{,}011(\text{Age}_i = x) + 0{,}009(\text{Female}_i = 1)(\text{Age}_i = x) = 0{,}354 + 0{,}02x$

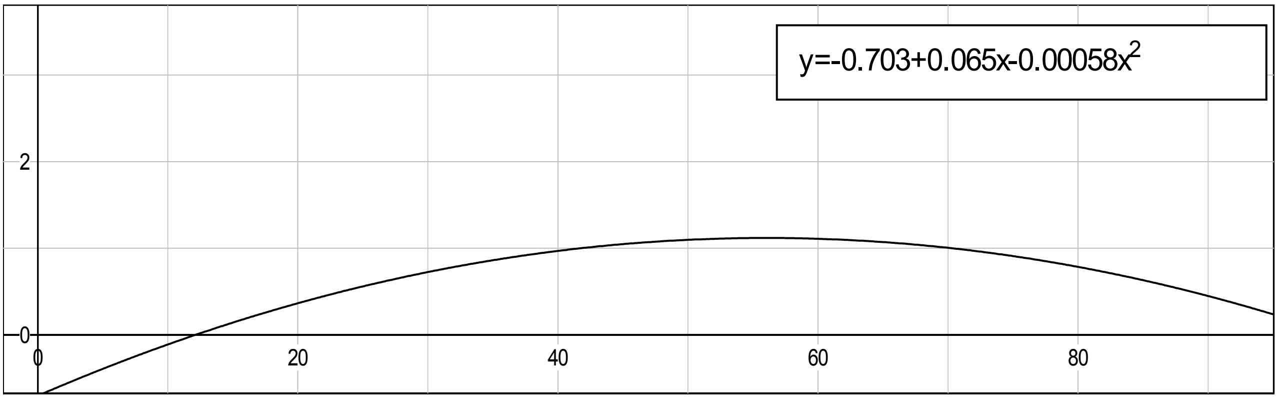
Since the coefficient of x is positive, we see that the probability increases with age. We can plot the predicted P = 1/(1+exp(-[0,354+0,02x])). For men the formula becomes P = 1/(1+exp(-[0,5+0,011x])).



If we would do the same for model 5 the problem become more complex since age is curvilinearly related to the logit. We have to determine what is more important: the age polynomial for men or the age polynomial for women?

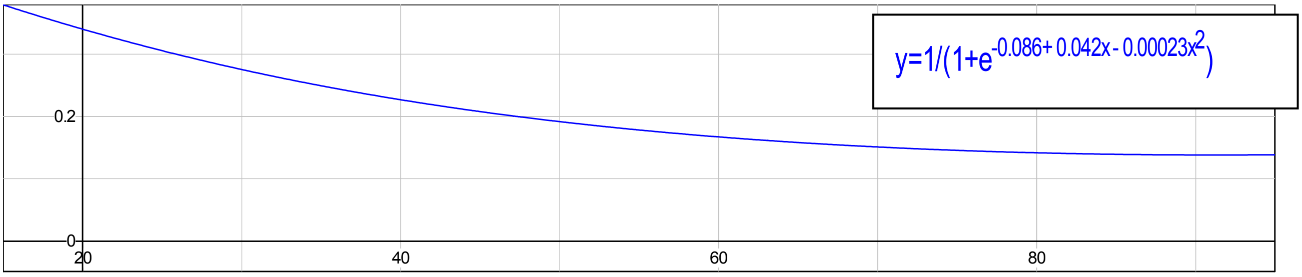
B	Variable	Minimum	Maximum	Value maximizing logit
1,063	Constant			
-,638	HedmarkOppland	,00	1,00	0
,184	SouthEast	,00	1,00	1
-,044	AgderRogaland	,00	1,00	0
-,294	WestNorway	,00	1,00	0
-,385	Troendelag	,00	1,00	0
,021	NorthNorway	,00	1,00	0
-,477	Voted	,00	1,00	0
,032	LeftRightScale	,00	10,00	10
-,703	Female	,00	1,00	1
-,090	Educ	5,00	20,00	5
-,100	FemaleEduc	,00	18,00	1*5
-,023	Age	17,00	93,00	X
,065	FemaleAge	,00	93,00	1*X
,00035	Age2	289,00	8649,00	X^2
-,00058	FemaleAge2	,00	8649,00	1* X^2

Women will maximize the logit if the difference between the two polynomials, -0,703(Female$_i$) + 0,065(FemaleAge$_i$) -0,00058(FemaleAge2$_i$), is larger than zero. From a plot of this polynomial we see that for reasonable values of age, 15< age < 95, women will have higher logit than men.



Hence the logit for the conditional effect plot that will maximize the probability according to age is
$L_i = 1{,}063 + 0{,}184(\text{SouthEast}_i = 1) + 0{,}032(\text{LeftRightScale}_i = 10)$
$-0{,}703\ (\text{Female}_i = 1) - 0{,}09(\text{Educ}_i = 5) - 0{,}1(\text{Female}_i = 1)(\text{Educ}_i = 5)$
$-0{,}023(\text{Age}_i = x) + 0{,}065(\text{Female}_i = 1)(\text{Age}_i = x) + 0{,}00035(\text{Age2}_i = x^2) -$
$0{,}00058(\text{Female}_i = 1)(\text{Age2}_i = x^2) = -\ 0{,}086 + 0{,}042x - 0{,}00023x^2$

We can plot the predicted $P = 1/(1+\exp(-[-\ 0{,}086 + 0{,}042x - 0{,}00023x^2]))$ and finds that the probability declines with age.



f) Find from model 5 an expression for the odds ratio for observing Y=1 on “Politicians interested in votes rather than peoples opinions” between groups of men with one year difference in age

The odds ratio between age groups with one year difference is defined as the ratio of exp(Logit(age=x+1))/exp(Logit(age=x)).

As argued above those parts of the logit equation that does not involve age can be seen as a constant called Const in the expression below. In model 5 four elements involve age:

	B
Age	-,023
FemaleAge	,065
Age2	,00035
FemaleAge2	-,00058

$$\mathrm{OR(age{+}1/age)}$$

$$= \frac{e^{\mathrm{Const}\ -0,023(\mathrm{Age}+1) + 0,065(\mathrm{Female}=1)(\mathrm{Age}+1) +0,00035(\mathrm{Age}+1)^2\ -0,00058(\mathrm{Female}=1)(\mathrm{Age}+1)^2}}{e^{\mathrm{Const}\ -0,023\mathrm{Age} + 0,065(\mathrm{Female}=1)\mathrm{Age} +0,00035\mathrm{Age}^2\ -0,00058(\mathrm{Female}=1)\mathrm{Age}^2}}$$

$$= \frac{e^{\mathrm{Const}\ -0,023\mathrm{Age}\ -0,023 + 0,065(\mathrm{Female}=1)\mathrm{Age}+0,065(\mathrm{Female}=1) +0,00035(\mathrm{Age}+1)^2\ -0,00058(\mathrm{Female}=1)(\mathrm{Age}+1)^2}}{e^{\mathrm{Const}\ -0,023\mathrm{Age} + 0,065(\mathrm{Female}=1)\mathrm{Age} +0,00035\mathrm{Age}^2\ -0,00058(\mathrm{Female}=1)\mathrm{Age}^2}}$$

$$= \frac{e^{-0,023\ +0,065(\mathrm{Female}=1)\ +0,00035(\mathrm{Age}+1)^2\ -0,00058(\mathrm{Female}=1)(\mathrm{Age}+1)^2}}{e^{+0,00035\mathrm{Age}^2\ -0,00058(\mathrm{Female}=1)\mathrm{Age}^2}}$$

The odds ratio for men and the odds ratio for women are then given as

$$\mathrm{OR}_{\mathrm{men}}(\mathrm{age}+1/\mathrm{age})$$

$$= \frac{e^{-0,023\ +0,00035(\mathrm{Age}+1)^2}}{e^{+0,00035\mathrm{Age}^2}} = \frac{e^{-0,023\ +0,00035(\mathrm{Age}^2\ +2\mathrm{Age}+1)}}{e^{+0,00035\mathrm{Age}^2}}$$

$$= \frac{e^{-0,023\ +0,00035(2\mathrm{Age}+1)}}{e^{0}} = e^{-0,02265\ +0,0007\mathrm{Age}}$$

$$\mathrm{OR}_{\mathrm{women}}(\mathrm{age}+1/\mathrm{age})$$

$$= \frac{e^{-0,023\ +0,065*1\ +0,00035(\mathrm{Age}+1)^2\ -0,00058*1*(\mathrm{Age}+1)^2}}{e^{+0,00035\mathrm{Age}^2\ -0,00058*1*\mathrm{Age}^2}}$$

$$= \frac{e^{0,042\ -0,00023(\mathrm{Age}+1)^2}}{e^{-0,00023\mathrm{Age}^2}} = \frac{e^{0,042\ -0,00023(\mathrm{Age}^2\ +2\mathrm{Age}+1)}}{e^{-0,00023\mathrm{Age}^2}}$$

$$\frac{e^{0,042\ -0,00023(2\mathrm{Age}+1)}}{e^{0}} = e^{0,04177\ -0,00046\mathrm{Age}}$$