Assignment 2 - Obama & Happiness - Discrete Choice Modelling

Marc Sparhuber

Table of contents

Data & Descriptives

The following analysis tests the hypothesis that individuals who were more unhappy in 2016, were more likely to have voted Obama in 2012. This is expected to be the case because in 2016 individuals who had voted Democrat in 2012 were now faced with Donald Trump as their future president, in some cases leading to protests (Domonoske, 2016). To test this hypothesis, a subset of data from the General Social Survey, conducted by the National Opinion Research at the University of Chicago and provided as part of the socviz package in R is used.

Table 1: Descriptives split by voting for Obama

		voted Obama				didn't vote Obama				combined			
		Mean	SD	N	Percent	Mean	SD	N	Percent	Mean	SD	N	Percent
Happiness	Very Happy			269	25.43			221	34.80			490	28.94
	Pretty Happy			622	58.79			355	55.91			977	57.71
	Not Too Happy			167	15.78			59	9.29			226	13.35
Bachelor or Graduate degrees of parents	0			787	74.39			475	74.80			1262	74.54
	1 or more			271	25.61			160	25.20			431	25.46
Sex	Male			420	39.70			310	48.82			730	43.12
	Female			638	60.30			325	51.18			963	56.88
Age		52.32	16.76	1058	100.00	56.00	16.31	635	100.00	53.70	16.68	1693	100.00
All				1058	100.00			635	100.00			1693	100.00

Comments: General Social Survey data from the socviz R package.

Table 1 displays descriptive statistics for all variables used for modelling, split by having voted Obama or not, as well as the entire sample. It shows number of respondents and percent within its group and variable for categorical variables and extends this by means and SDs for continuous variables. Of the 1693 individuals included in the analyses, 1058 (62.49%) voted for Obama, while the 635 (37.51%) did not. It is important to note that this indicates a slightly skewed sample, as Obama received 51.1% of the popular vote in 2012 (Federal Election Commission, 2013). As the main independent variable, Happiness is split into the categories 'Not Too Happy', 'Pretty Happy' and, 'Very Happy'. Across the entire sample, more than half of the respondents reported being 'Pretty Happy' (57.71%), 28.94% reported being 'Very Happy', while 13.35% considered themselves 'Not Too Happy'. Relevant to the hypothesis, while only 9.29% of those who did not vote for Obama reported being 'Not Too Happy', 15.78% of those who voted Obama did. While the responses for being 'Pretty Happy' were similar across respondents who had voted Obama or not - 58.79% and 55.91%, respectively - noticeably fewer of the Obama voters considered themselves 'Very Happy' (25.43%) than the others (34.80%).

Further categorical variables describing the sample include whether the respondent had at least one parent with a higher education degree, sex, and age. Interestingly, there seems to be no connection between having voted Obama and higher education, with 75% of this sample having no college educated parent and this percentage

not varying across having voted Obama or not. Looking at the sex variable, although 56.88% of respondents were female, voting seemingly split along gender lines with women outweighing men among Obama voters with 60.30% and men being more represented among those who did not vote for Obama with 48.82%. Regarding the age of respondents, the sample mean is 54 years. This varies only slightly across the two conditions, with Obama voters being slightly younger average (52) than non-Obama voters (56). The standard deviation stays at at around 16 in all conditions.

Model Estimaton & Odd Ratios

Table 2 shows the results of three logistic probability models with voting for Obama as their dependent variable. Each model contains the full sample of 1693 observations and presents the coefficients as odds ratios with corresponding confidence intervals in square brackets. Model 1 uses Happiness as its sole independent variable, with 'Not Too Happy' as its reference category. Compared to this reference category both 'Very Happy' and to a lesser extent 'Pretty Happy' have significant negative effects (due to their odds ratios being below 1) on voting for Obama which does not change across all three models. This can be interpreted as the odds of voting for Obama being multiplied by 0.619 (which means that they decrease) in the case of being 'Pretty Happy' and 0.430 for being 'Very Happy'. Interestingly, whether one of the parent's has a college degree does not seem to have an effect at all, being insignificant in Model 2 and 3 and with its confidence intervals crossing the value of OR = 1, as can be seen in Figure 1.

Table 2: Voting for Obama. Logistic probability models

|	Model 1	Model 2	Model 3
Pretty Happy	0.619**	0.615**	0.614**
	[0.445, 0.852]	[0.442, 0.847]	[0.440, 0.849]
Very Happy	0.430***	0.427***	0.431***
	[0.303, 0.605]	[0.300, 0.602]	[0.302, 0.609]
At least one parent has degree		1.061	0.924
		[0.845, 1.335]	[0.729, 1.174]
Age			0.985***
			[0.979, 0.992]
Female			1.483***
			[1.212, 1.815]
Num.Obs.	1693	1693	1693
AIC	2220.6	2222.3	2192.1
BIC	2236.9	2244.1	2224.7
Log.Lik.	−1107.301	−1107.172	−1090.037

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Source: GSS data from the socviz R package.

Comments: The reference cateogory for happy is 'Not Too Happy'.

Model 3 includes age and sex as well. Both of these are significant, but the odds ratio for age is close to zero, indicating that is has no effect on voting for Obama, as well as a very narrow confidence interval as can be seen in Figure 1. Regarding sex, the odds of voting for Obama are 1.483 times as large for women than for men, although this odds ratio has a relatively large confidence interval as displayed in Figure 1, meaning that generalizing this finding about sex beyond this sample may not be as smooth as for age. Further, the AIC, BIC, and the Log. Likelihood for each model are computed. In Figure 1, the red line at odds ratio = 1 indicates no association between the odds ratios of the corresponding predictors and their confidence intervals above the line indicate a positive effect and vice versa.

Figure 1

Voting for Obama: Logistic Regression

GSS data from the socviz R package.
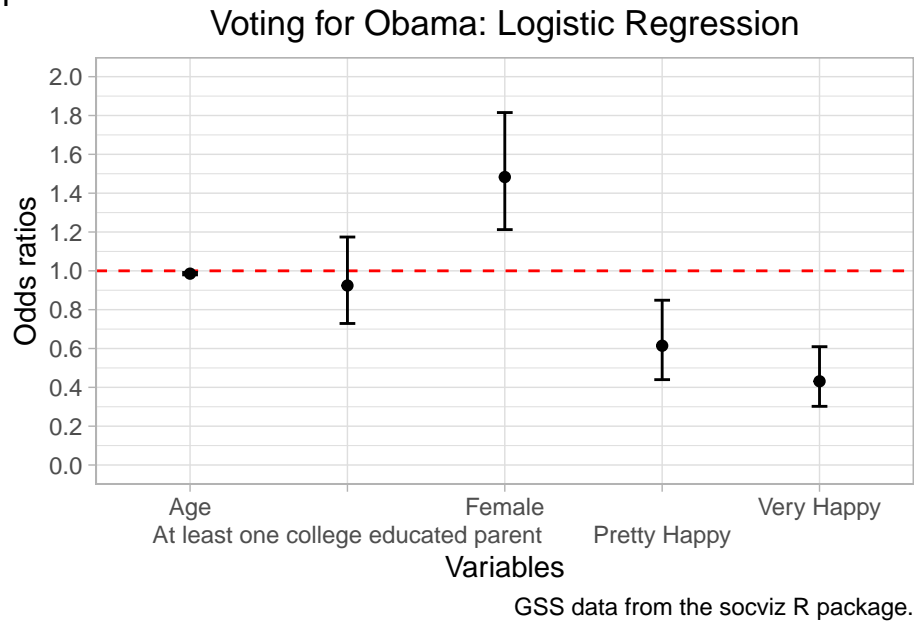
Predicted Probabilities

Figure 2 presents the predicted probabilities of voting for Obama in 2012 for women with at least one college educated parent based on the previously run logistic regression model 3, including the predictors age, sex, parental educational attainment and happiness. On the x-axis, the range of all ages of all female participants is used, while the y-axis shows the predicted probability of voting Obama. These probabilities are split by Happiness. Across all three categories of happiness a gradual decrease in predicted voting for Obama can be seen. This implies that the older a woman with at least one college educated parent, the lower the predicted probability of her voting for Obama. Happiness seems to also affect this predicted probability, as the unhappier an individual, the greater the predicted probability of having voted Obama. This is in line with the previously posed hypothesis.

5

Figure 2
Probability of Obama vote for women with at least one college educated parent

GSS data from the socviz R package.
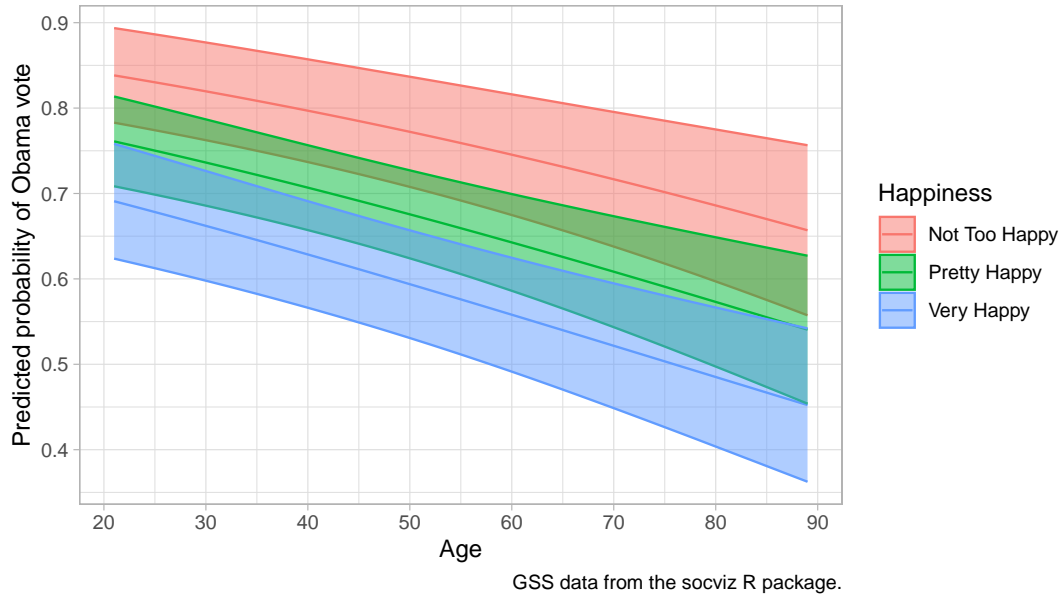
Model Fit

Table 3 presents the model fit statistics for all three models. These models are nested within each other, looking at the graph from the left to the right, as they all contain a subset of the parameters used in the model to their right. While Nagelkerke's pseudo R^2 is similar in the first two models, this changes when introducing age and sex in model 3. Here, Nagelkerke's pseudo R^2 rises from around 0.02 to 0.047, which indicates that the third model performs better than the previous two. The share of correct predictions stays similar across all three models, demonstrating that the added predictors do not enchance model 2 and 3's accuracy. Calculating the likelihood ratio of model 1 to 2 and 2 to 3 it is evident that while adding parent's educational attainment to the model does not increase its fit (and results in an insignificant result), adding age and sex in model 3 does so, significantly.

Table 3: Voting for Obama: Model fit statistics

	Model 1	Model 2	Model 3
Nagelkerke's pseudo-R2	0.02044	0.02064	0.04754
Share of correct predictions	0.62493	0.62493	0.62788
Likelihood Ratio		0.25893	34.26947
Likelihood Ratio (p-values)		0.61086	3.61808e-08

Note: A prediction is considered correct when its probability is greater than 0.5. The Likelihood ratio is always calculated with the nested model to the left. Data from GSS 2016.

Conclusion

Overall it can be said that the initially posed hypothesis is supported by the above results, as being less happy in 2016 predicts having voted for Obama in 2012, as can be seen in Figures 1 and 2 and Table 1 and 2. Interestingly, while age by itself seems to not affect the odds of voting for Obama for the entire population, it does increase the predicted probability in a specific subset, which may warrant further research.

References

Domonoske, C. (2016). Across The Country, Crowds March In Protest Against Trump's Victory. In *NPR*. https://www.npr.org/sections/thetwo-way/2016/11/09/501513889/anti-trump-protests-break-out-in-cities-across-the-country

Federal Election Commission. (2013). *Federal Elections 2012*. https://www.fec.gov/resources/cms-content/documents/federalelections2012.pdf#page=11