2024

ST211 Individual Project

CANDIDATE NUMBER: 28094

WORD COUNT: 1499

Introduction

This report analyses the 2012 Northern Ireland Life and Times Survey, a Dataset focusing on people's views towards LGBT people in the UK. The data were collected through face-to-face interviews with adults in the UK, and contain many different factors of interest, such as respondents' age, gender, social class and more.

I analyse it to answer the following question; 'what factors affect people's views on the rights of same sex marriages?' This question is particularly relevant, as same sex marriages do not possess the same rights as traditional marriages in many parts of the world, and it has also been relatively recent since same sex marriages have been allowed in England[1]. Thus, investigating the factors that can cause certain views to emerge is worth doing.

The analysis finds that variables such as people's age, gender, frequency of religious service attendance and happiness are significant predictors of responses to the question of same sex marriage rights.

Exploratory Data Analysis

The initial stages of the investigation focused on understanding the dynamics between various predictors and their impact on responses to the question 'Should same sex marriages have the same rights as traditional marriages?' (coded as the variable ssexmarr). This was carried out through a variety of plots.

Addressing Missing Values:

My analysis removed missing values for most predictors and generally viewed the data as a complete case study. This was because, in most cases, missing values seemed to occur randomly, and so including them as a level would likely just be measuring noise, leading to overfitting in the model. The exceptions were variables like famrelig, chattnd2, and polpart2 where I felt missing values had real significance. For example, the only possible responses to the 'famrelig' variable, which refers to the question 'In what religion were you brought up?', were 'Catholic', 'Protestant' or 'No religion'. This clearly excluded any other possible religion[2] so I took 'missing' as possibly referring to another religion. Finally, I removed predictors with more than 20% of missing values, such as rsuper, rsect and persinc2, as I could not find a consistent reason for missing data and felt removing that many missing rows would make the model weaker.

[1] In 2013

[2] E.g. Islam, Hinduism, Buddhism

Predictor Evaluation:

Most predictors appeared somewhat significant. Below I include a few variables that did seem different from the norm, as well as variables that caused me to alter my model:

umineth
Ethnic group did not appear significant.

Famrelig and religcat
Both variables seemed very similar and had a 78% correlation with one another, so I decided to only use one in the future model, to prevent multicollinearity.

persinc2
Personal Income appeared insignificant in the plots, affirming my earlier decision of removing it due to its missing values.

I did not remove any predictors based on apparent significance however, as I would do this in the Regression stage.

Merging Levels:

I merged multiple levels to improve interpretability. For example, several levels in chattnd2 had very similar effects on the outcome and could be grouped into one 'less frequent' level.[3]

[3]A full table of variables which had levels merged is provided at the back of the report.

From Initial to Final Model

After the EDA, I began the logistic regression analysis. My first model was a logistic regression with every predictor shortlisted after the EDA included. It found that the majority of predictors were insignificant at the 5% level. I only included significant predictors in the next model. Additionally, I added back in missing rows which I had removed for insignificant predictors. In the subsequent model, the difference between null and residual deviance was higher than the first (443.7 compared to 406.0).

However, though most predictors were significant at the 5% level in model 2, the 'Highest Qualification' predictor (highqual) was not, though only barely. I decided to run a few more diagnostics before deciding whether to remove it. Firstly, I tried a model without highqual, which ended up having a lower difference between null and residual deviance. Then I created a classification table, to measure how often my model correctly predicted the outcome of ssexmarr. It correctly predicted when people would be supportive of gay marriage rights 92% of the time, however, only predicted people being against gay marriage rights 62% of the time correctly. Without highqual, this improved to 63% of the time. This led me to exclude highqual for my third model, due to its higher p-value and lower predictive power. Though this did lead to a lower difference between null and residual deviance (438.6 compared to 443.7), I felt the cost was worth the benefit.

My final model included an interaction (and I also centred the age predictor to improve interpretability). I focused on the age variable, as it was highly significant and made the most sense theoretically to have an interaction. I found a significant interaction[4] between age and the response to the question of 'how would you feel if your child was gay/lesbian?' I included this in my final model, which ended up having the highest difference between null and residual deviance of all previous models (456.0 compared to the next highest 443.7), and also a slightly better classification table, correctly predicting 64% of people as being against gay marriage compared to 63% in the previous model.[5]

[4] At the 5% level

[5] Though this was a very minor difference

Results

The model from the display function is shown below.

Predictor	coef.est	coef.se
(Intercept)	-3.27	1.02
cent.age	0.05	0.01
rsexMale	0.61	0.19
chattnd2Missing	-0.39	0.28
chattnd2Never	-0.64	0.32
chattnd2Once a week	0.12	0.25
chattnd2Several times a week	0.89	0.39
ruhappyHappy	1.00	1.00
ruhappyNot at all happy	4.10	1.52
ruhappyNot very happy	1.91	1.05
glchildNeither comfortable nor uncomfortable	1.24	0.25
glchildUncomfortable	0.94	0.40
glsocdist1	0.76	0.33
glsocdist2	0.91	0.43
glsocdist3	0.50	0.47
glsocdist4	2.25	0.67
glsocdist5	1.65	0.63
glsocdist6	18.16	918.80
glsocdist7	1.89	1.22
glsocdist8	0.54	0.89
glsocdist9	18.41	1289.90
glsocdist10	18.24	1238.44
glsocdist11	2.86	0.73
glbornChoose to be gay/lesbian	1.17	0.23
cent.age:glchildNeither comfortable nor uncomfortable	0.00	0.01
cent.age:glchildUncomfortable	-0.06	0.02

Particular Variables:[6]

Age:

Increases in age increase the probability of not supporting same sex marriage rights, generally, however, the interaction term can reverse the direction.

[6] More detailed examples in the Interpretation section

Gender:

Men are more likely to not support same sex marriage rights.[7]

Religious Services:

Attending religious services more frequently increases the probability of being against same sex marriage rights.

Happiness:

Greater unhappiness increases the probability of being against same sex marriage rights.

Comfort with LGBT people:

In general, people who feel uncomfortable about LGBT people in multiple domains[8] are more likely to be against same sex marriage rights.

[7] Compared to women

[8] E.g. when asked about having a gay child, and also questioned on the number of social scenarios in which they'd feel uncomfortable about an LGBT person being near

Comments about the Data/Analysis

It should be noted that the data was gathered in 2012 and views about LGBT people in the UK would likely have shifted since then. For example, there have been demographic changes in the UK due to immigration, whilst the newest cohort of adults may have largely different views from previous generations. Thus, certain predictors may be less relevant in the modern day and the predictive power of the model is likely reduced.

Additionally, people may have lied about certain variables. For example, some may not have felt comfortable discussing their sexual orientation with an interviewer or may have feared reproach if their views on same sex marriage differed from the norm.

Furthermore, the ‘glsocdist’ variable in my model is not very clear. The data dictionary describes it as the ‘number of situations in which a respondent felt uncomfortable if a person was gay/lesbian’ but these situations are not described.

Finally, although I assumed missing values in this dataset were mostly missing randomly, missing values did seem to have different effects on the outcome in my exploratory plots. Thus, there could have been a systemic reason for missing data that I did not consider.

Interpretation

The model I've built tries to find out what makes it more or less likely people believe same sex marriages should have the same rights as traditional marriages.

It finds that there are a few important variables. These are age, gender, frequency of attending religious services, as well as how comfortable people feel about LGBT people (which can be broken down into factors like how they'd feel about a gay child, for example).

One of the most important factors is age. Take two men that both do not attend religious services, are happy, and feel largely comfortable with gay people, but one man is 20 and the other is 50. The odds of this 20-year-old supporting gay marriage are 77 to 1, whilst the odds of the 50-year-old supporting gay marriage decrease to 17 to 1. However, if we look at two people who are uncomfortable with gay people, age actually has the opposite effect. For example, once again, take two men, one 50 and the other 20, but this time both of them feel uncomfortable about the idea of having a gay child. This time the 20-year-old has only a 5 to 1 chance of being supportive of gay marriage, whilst the 50-year-old is predicted to have a 7 to 1 chance.

Gender is also important. For a comparable 30-year-old man and woman that both feel neither comfortable nor uncomfortable about gay people, the woman has 9:1 odds of being supportive of gay marriage whilst the man has a smaller, 5:1 odds.

However, you should view this model as far from perfect. When someone actually *is* against same sex marriage rights the model only correctly predicts that 64% of the time. For any thinking of using the model for official decision making[9], I would also add that it does not give any input on what actually causes people to support or be against same sex marriage and there could be important variable our data has not considered that are the *actual* main reasons people answer a certain way.

[9] (Please use Dr Geneletti's model instead)

Appendix:

Predictor	Shortlisted after EDA?	Merged Levels?
househld	Y	N
rage	Y	N
rsex	Y	N
Rmarstat	Y	N
Livearea	Y	N
Hincpast	Y	N
Intwww	Y	N
Umineth	Y	N
Eqnow3	Y	N
Eqnow7	Y	N
Eqnow9	Y	N
Eqnow11	Y	N
Tenshort	Y	Y
Highqual	Y	Y
Tea	Y	N
Work	Y	N
Rsuper	N	N
Rsect	N	N
Tunionsa	Y	N
ansseca	Y	N
Religcat	N	N
Famrelig	Y	N
Chattnd2	Y	Y
Carehome	Y	N
Anyhcond	Y	N
Persinc2	N	N
Orient	Y	N
Polpart2	Y	N
Ruhappy	Y	Y
Healthyr	Y	Y
Uprejgay	Y	N
Glchild	Y	Y
Glsocdist	Y	N
Glvis	Y	N
Glborn	Y	N
Knowgl	Y	N
Knowtg	Y	N