ST211 Individual Project Report

Candidate Number : 22757

WORD COUNT

(Excl. cover page, footnotes, titles, tables & appendix)

1499

I. INTRODUCTION

In 1967, the UK decriminalised private homosexual acts.

Yet, it took four more decades before same-sex marriage was legalised. This logistic regression analysis aims to investigate the various factors influencing people's stance on supporting equal same-sex marriage rights to traditional marriages. Understanding these causes will enable social activists and policy makers to better navigate the ongoing challenges and opportunities within the movement for marriage equality.

Utilising a focused analysis of 1064 responses from the 2012 NILT dataset, this report will elaborate how the significant influence of demographic factors, such as age and religion, and social attitudes[1], particularly prejudices against homosexuals, shapes an individual's support for equal rights in same-sex marriage.

[1] Curtice, S., Ratti, V., Montagu, I., & Deeming, C. (2023). *A new generational divide? The age gap in British political attitudes*. National Centre for Social Research. https://natcen.ac.uk/publications/bsa-40-age-differences

II. EXPLORATORY DATA ANALYSIS

2.1 Approach

To understand the data, plots were run against the outcome variable for each predictor. Specifically:

- for categorical predictors, fill plots allowed examination of within group percentages.
- for continuous predictors, flipped box plots & logistic curves with scatter plots were used.

2.2 Sanity Checks

Bias (see comments[2]), dependencies (see appendix[2]).

2.2 Missing Data

Upon close examination of the data through plots and the data dictionary, some variables with missing data are observed to have an intuitive explanation. For instance, *chattnd2* has substantial missing data because this question was skipped for non-religious people. Similarly, for *work, rsuper* (occupational questions), the missing data likely pertains to the unemployed, as the question was not applicable to them.

As such, for the initial model, these missing data were included (encoded to "missing"), and later, effects of excluding them are examined in the complete case analysis. Also, predictors with low percentages of missing data (<4%) were removed due to low sample sizes, which added noise and have limited explanatory power; for other predictors, the missing data was encoded to missing, and some were merged with other levels.

2.4 Thoughts on Plots

Below are some promising predictors based on the plot observations. If the final model omits most of them, it could indicate that something went wrong.

Predictor	Expect to be Significant?	Thoughts
rage	Yes	Logistic curve looks quite pronounced, older people seem less supportive.
rsex	Yes	Females seem more supportive.
religcat	Yes	Catholic and non-religious people are much more supportive than Protestants.
uprejgay	Yes	Distinct differences across levels.

Similarly, for predictors expected to be very insignificant:

Predictor	Expect to be Significant?	Thoughts
hincpast	No	Consistent within group percentages across levels.
umineth	No	
tunionsa	No	
carehome	No	Consistent within group percentages between Yes and No.

[2] Elaborated in the comment section.
[3] Not significant towards analysis, discussed in model building; methodology elaborated in the appendix section.

TABLE 2.1: Exploratory Data Analysis Summary

Predictor	Expect to be Significant?	Initial Thoughts	% Missing	Merge Levels?	
household	Yes	Slightly supportive with more siblings.		Yes	
rage	Yes	Logistic curve looks quite pronounced, older less supportive.	0.19%	n/a	
rsex	Yes	Female seems more supportive.		No	1
rmarstat	Yes	Slightly different level of supportiveness across different groups.	0.19%	Yes	
livearea	Maybe	Lower years, slightly more supportive, not as pronounced as rage.		n/a	
hincpast	No	Consistent within group percentages across levels.	4.51%	No	
intwww	Yes	Individuals with internet access seem more supportive.		No	
umineth	No	Consistent within group percentages across levels.	0.09%	No	
eqnow3	Yes	Those answered yes seem more supportive.	0.38%	No	
eqnow7	Yes		0.38%	No	
eqnow8	Yes		0.38%	No	
eqnow9	Yes		0.38%	No	
eqnow11	Yes		0.38%	No	
tenshort	Yes	Houseownership group less supportive.	0.56%	Yes	
highqual	Yes	Lower qualification less supportive.		Yes	
tea	Yes	Completed school early not so supportive.	0.94%	Yes	
work	Maybe	Self-employed or other not so supportive, lots of missing data.	14.19%	No	2
rsuper	Maybe	Difference between those who answered and those who did not.	25.19%	No	2
rsect	Maybe	Some are supportive, but many missing data, could affect statistical significance.	26.50%	Yes	
tunionsa	No	Consistent within group percentages across levels.	1.22%	No	
ansseca	Yes	Similar observation to work.		No	
religcat	Yes	Catholic and non religious moderately more supportive.	1.79%	No	
famrelig	Yes	Similar to religcat. Perhaps correlated, further investigation needed.	2.63%	No	
chattnd2	Yes	Less religious activity, more supportive, some missing data.	18.42%	Yes	3
carehome	No	Consistent within group percentages.		No	
anyhcond	Yes	No sickness more supportive.	0.28%	No	
persinc2	No	Lots of missing data.	23.12%	n/a	
orient	Maybe	Sample size for non-heterosexual groups very small.	0.38%	No	
polpart2	Maybe	Lots of missing data, hard to determine.	10.90%	Yes	
ruhappy	Yes	Happy people more supportive		Yes	
healthyr	Yes	Difference in within group percentages across levels, has many levels.	0.66%	Yes	
uprejgay	Very	Very distinct differences across.	1.69%	Yes	
glchild	Yes	Fairly distinct differences across levels.	1.50%	Yes	
glsocdist	Maybe	Range of levels, more supportive with decreasing situations.		Yes	
glvis	Yes	Difference between 0 and the rest in within group percentages.		Yes	
glborn	Yes	First statement less supportive.	11.37%	No	
knowgl	Maybe	Difference between know and dont know, some missing data.	8.27%	No	4
knowtg	Maybe		11.56%	No	4

Note:
1 | Might contain some bias: 43% male, 67% female.
2 | Missing data is likely corresponds to response by unemployed person.
3 | Question is skipped if not religious.
4 | Missing data could be people who are unsure.

2.5 Merging Levels & Resetting Baseline Level

Based on insights from plot analyses, some predictors contained many levels with similar outcomes, and some levels had low sample sizes, which could have a small effect size. Thus, levels for some predictors were consolidated through observations from plots and with consideration from a simple logistic regression.

Detailed information of merged levels are included in the appendix. Furthermore, to ensure higher statistical power, the baseline was set using the group that had the most observations.

III. FROM INITIAL TO FINAL MODEL

3.1 Model 1 (Full Model)

The full model outperforms the null model, as shown by a decrease in deviance of 438.29. It is significant by Chi-squared test, and yielded several meaningful predictors, suggesting that the logistic regression model is potentially a good fit.

3.2 Model 2 (Backwards Elimination)

The full model contained many insignificant predictors, which made the model complex. To streamline it, a backward elimination approach was employed. Predictors were iteratively removed based on ANOVA Chi-squared and VIF (<10) assessments , checking insignificance (at 5%) and multicollinearity respectively. This resulted in a refined model with 13 significant predictors.

Model 2 has a decrease in deviance from the full model of 95.55, and a slight drop of 3% in overall prediction accuracy. Moreover, the full model displays a better fit through a Chi-square test at a 5% significance level. That said, this reduced model is leaner and more interpretable.

3.3 Dependency Analysis

At EDA, *religcat* and *famrelig* were suspected to be correlated, so a separate analysis was run, but there were no significant differences.

3.4 Outlier Analysis

An attempt to find outliers was made, but no datapoint violated all 3 criterias: DFFITS, Leverage, Cooks.

3.5 Model 3 (Transformation)

An attempt to center *rage* (age) predictors was made to improve interpretability as the new intercept will represent the estimated log odds of the outcome at the mean age.

3.6 Model 4 (Interaction)

Interactions were applied between various significant variables. One pair of significant interacting variables was found (*religcat* * glsocdist), but the individual levels were not significant. Perhaps a bigger sample size is needed to capture such relationships. Moreover, this interaction did not seem particularly intuitive, so this model is not pursued further.

3.7 Model 5 (Complete Case Analysis)

In the complete case analysis, the variable *rsect* (sector of employment) emerged as significant. It is important to note, however, that this reduced model excluded half of the original dataset, which is a significant reduction, consequently, this result could be biased.

TABLE 3.1: Significance Level of Predictors via Chi-squared Test (ANOVA)

Key:
. means significant at 10% level
* means significant at 5% level
** means significant at 1% level
*** means significant at 0.01% level

Predictor	**Model 1** Full Model	**Model 2** Backwards Elimination	**Model 3** Centred Model 2	**Model 4** Interaction	**Model 5** Complete Case
rage	***	***	***	***	***
uprejgay	***	***	***	***	***
glchild	***	***	***	***	***
rsex	***	***	***	***	**
religcat	***	***	***	***	*
ruhappy	***	***	***	***	(removed)
glsocdist	***	***	***	***	(removed)
househld	***	***	***	(removed)	(removed)
eqnow11	**	**	**	(removed)	***
eqnow3	*	**	**	(removed)	(removed)
anyhcond	**	*	*	(removed)	**
intwww	*	*	*	(removed)	(removed)
eqnow8	*	*	*	(removed)	(removed)
rsect		(removed)	(removed)	(removed)	**
ansseca	*	(removed)	(removed)	(removed)	(removed)
tenshort	.	(removed)	(removed)	(removed)	(removed)
highqual	.	(removed)	(removed)	(removed)	(removed)
orient	.	(removed)	(removed)	(removed)	(removed)
glvis	.	(removed)	(removed)	(removed)	(removed)
glborn	.	(removed)	(removed)	(removed)	(removed)
chattnd2	*	(removed)	(removed)	.	(removed)
religcat:glsocdist	n/a	n/a	n/a	*	n/a
Δ Deviance	-438.29	-342.74	-342.74	-334.8	-126.44
Null Deviance	999.58	999.58	999.58	999.58	418.85
Residual Deviance	561.29	675.54	675.54	664.78	292.41
Correct Prediction (%)	0.80	0.77	0.77	0.75	0.73
Correct for support	0.91	0.92	0.92	0.91	0.88
Correct for don't support	0.68	0.62	0.62	0.58	0.57

Note: Variables that were removed from model 2 to 5 are omitted from this table. These variables are included in the appendix.

Model 3 is chosen as the final model. Despite the fact that this model displayed an increase in deviance (of 95.55) from the full model, it has lesser predictors and therefore is easier to interpret. This model is statistically significant from the null model by Chi-squared test, and has an overall correctness of 77%[4], which is reasonably good; however, future studies could benefit from exploring non-linear models.

[4] Further discussion on this score in the comments section.

IV. RESULTS

Here are the most significant *categorical predictors* identified by Model 3:

TABLE 4.1: Top 5 Most Significant Variables for Model 3

Variable	Level	Baseline Level	Coefficient	Odds	P-value
uprejgay	A_little	Not	1.07	2.92	1.49e^(5)
glchild	Indifferent	Very_comfortable	1.20	3.32	5.83e^(5)
glchild	Fairly_uncomfortable	Very_comfortable	1.65	5.23	2.20e^(5)
religcat	Catholic	Protestant	-0.58	0.56	0.00891
ruhappy	Unhappy	Happy_or_indifferent	1.07	2.91	0.00388

Prejudice & Discomfort with Child being Homosexual

Unsurprisingly, for people with slight prejudice against gay people, the odds of not supporting equal same-sax marriage rights is 2.92 times that of those who are not prejudiced. Similarly, if an individual is either indifferent to or uncomfortable with their child being homosexual, their odds of not supporting is about 1.20 and 1.65 times that of those who are comfortable with their child's sexuality respectively.

Catholic vs Protestant

An interesting predictor emerged in the model is the role of religious affiliation, indicating that Catholics are less likely to oppose equal rights marriages compared to Protestants, with odds of 0.58 to 1, when other variables are set to the baseline. This could suggest that Catholics are more accepting of homosexuality, one study's findings aligns with this.[5]

Happiness

The model suggests a notable correlation between happiness levels and support for same-sex marriage. Specifically, individuals who are generally unhappy have an odds of 2.91 times to those are happy or indifferent when it comes to supporting same-sex marriage rights. This association may suggest that a more pessimistic worldview could influence openness to changes in social norms.

[5] Pew Research Center. (2003, November 18). Religious beliefs underpin opposition to homosexuality. https://www.pewresearch.org/politics/2003/11/18/religious-beliefs-underpin-opposition-to-homosexuality/

Notably, the most significant predictor of this model is rage (age), a *continuous variable*:

Age

Model 3 predicts that an individual aged 18 has 0.1358 higher in probability of supporting equal same-sex marriage rights than a 60-year-old, meanwhile an individual aged 48 is 0.0527 higher in probability of supporting these rights than a 60-year-old. These results indicate the presence of progressively liberal views on same-sex marriage among younger generations.

Also, some risk ratios are calculated and included in the table below:

TABLE 4.2: Probabilities of Not Supporting by Age

Age	**Probability of *NOT* supporting** *(inverse logit of linear predictor)*	**Risk Ratio** *(between ages 60 vs. each age group)* [6]
18	0.0643	3.04 [7]
25	0.0786	2.47
35	0.1039	1.82
48 (mean)	0.1475	1.21
60	0.2002	N/A

[6] Risk ratio := P(*Not supporting at Age 60*) / P(*Not supporting at Age X)*

[7] In layman terms, a risk ratio of 3.04 means an individual at 60 is around 3 times as likely to oppose these rights than at 18.

TABLE 4.4: Model 3 Summary Table

Variable	**Level**	**Coefficient** (2 d.p.)	**Odds Ratio** (2 d.p.)	**P-value**	
Intercept		-1.75		7.77E-15	***
rage		0.03		8.74E-06	***
uprejgay	A_little	1.07	2.92	1.49E-05	***
glchild	Indifferent	1.20	3.32	5.83E-05	***
glchild	Fairly_uncomfortable	1.65	5.23	2.20E-05	***
religcat	Catholic	-0.58	0.56	0.00891	**
ruhappy	Unhappy	1.07	2.91	0.00388	**
glsocdist	More_than_5	1.89	6.65	0.00105	**
religcat	No_religion	-0.66	0.52	0.03665	*
glsocdist	3_to_5	1.14	3.14	0.02401	*
glchild	Very_uncomfortable	1.09	2.98	0.03747	*
rsex	Female	0.38	1.46	0.06449	.
eqnow8	Yes	-0.87	0.42	0.08509	.
househld	2	0.21	1.24	0.40539	
househld1	1	0.27	1.30	0.35731	
intwww	No	0.05	1.05	0.85988	
eqnow3	Yes	0.09	1.09	0.72715	
eqnow11	Yes	-0.30	0.74	0.25054	
anyhcond	Yes	0.17	1.18	0.48232	
uprejgay	Very	0.80	2.23	0.1124	
glchild	Fairly_comfortable	0.28	1.33	0.26158	

V. COMMENTS

Source of Bias

As with any study, it is essential to consider potential biases. Age, a key predictor in our analysis, showed a distribution similar to that of the general population, although (caveat) a thorough statistical validation was not done.

Moreover, there seems to be slightly more females (56.7%) than males in the data of this study when compared to the UK governmental data on proportion of female (51%)[8] to male. So a t-test was conducted at 5%. The results[9] did not indicate that the sample distribution deviated from the population distribution.

Outcome Data Imbalance

The final model achieves 77% accuracy, predicting supportive outcomes with 91% accuracy and non-supportive outcomes with 62% accuracy, indicating a better performance in predicting supportive outcomes. This disparity may be due to the dataset's predominance of supportive outcomes (64%). Furthermore, the consistency of these results across all models suggests a logistic regression model may not fully capture certain complexities within the data. Therefore, these results should be interpreted with caution.

TABLE 4.3: Confusion Matrix

	Observed : Supportive	Observed : Not supportive
Predicted : Supportive	451	103
Predicted : Not supportive	44	170
% Correct	91%	62%

Impact Missing Data

Although most predictors remain consistent in the complete case analysis, new significant variables have emerged. However, the exclusion of certain data, which might introduce biases, makes it challenging to endorse the model. That said, the absence of such gaps could lead to additional significant variables like *rsect*, as suggested by the analysis plots. Further studies may help verify its significance.

Furthermore, some variables such as chattnd2, work, and rsuper likely have substantial missing data due to limited response options available in the questionnaire. But according to the data dictionary, there are some instances of actual missing data. So, future research should focus on enhancing question options, which could enable better differentiation of missing and not applicable responses for interpretability purposes.

[8] GOV.UK. (31 March 2023) Male and female populations. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/male-and-female-populations/latest/

[9] Check the appendix for code and results.

VI. LAYMAN REPORT

In 1967, the UK made a significant legal stride by decriminalising private homosexual acts. Building on this historical context, this study examines the 2012 NILT dataset to dissect factors evolving public support for equal rights between gay and traditional marriages within individuals.

This analysis reveals a pronounced generational divide: young adults display more progressive attitudes compared to older people. Specifically, a 60-year-old is on average 3 times more likely to oppose marriage equality than an 18-year-old. This trend underscores a broader societal shift towards accepting diverse family and relationship structures with younger generations.[10]

As we dive deeper into the sociocultural influences on these views, significant differences emerge across religious affiliations. Catholics, for example, tend to support same-sex marriage more than Protestants, which may reflect a difference in the community's attitudes around homosexuality. Notably, according to our analysis, a 20-year-old female Catholic is found to be 6 times more likely to support equal rights compared to a 60-year-old Protestant male.

Furthermore, the analysis identifies ingrained prejudices, even at the slightest level, are much less likely to support same-sex marriage. This opposition is even more pronounced among those uncomfortable with the possibility of their children being homosexual: a 40-year-old man, who is uncomfortable with his child potentially being homosexual is 5 times more likely to oppose same-sex marriage rights than the average mother who is more accepting.

These findings offer a revealing glimpse into the factors influencing attitudes towards same-sex marriage, yet it is crucial to acknowledge the limitations of this study, particularly some missing data in a few survey questions. Such gaps in the data suggests that there may be other significant influences not captured in our analysis, and future research could address these deficiencies, allowing for a more comprehensive understanding of the evolving public support for marriage equality.

[10] Curtice, S., Ratti, V., Montagu, I., & Deeming, C. (2023). *A new generational divide? The age gap in British political attitudes*. National Centre for Social Research. https://natcen.ac.uk/publications/bsa-40-age-differences

VII. APPENDIX

7.1 Dependency Analysis during EDA

`cor(data)` was used to evaluate correlation between variables, in particular, *famrelig* and *religcat* were found to be highly correlated (0.8), thus this led to a dependency analysis when deriving the final model. As mentioned in 3.3, this did not yield any significant findings.

7.2 Bias Check

The proportion of females in this dataset was tested against the hypothesised true population proportion of 51% using a proportion test without continuity correction. The test yielded a p-value of 0.57, well above the significance threshold of 0.05, indicating that the observed proportion of females in our sample does not significantly differ from the expected population proportion. Therefore, there is no statistical evidence to suggest a deviation from the gender distribution of the general population, affirming the representativeness of our sample in terms of gender balance.

```
prop.test(546, 546+416, p = 0.51, correct = FALSE)
```

7.3 Omitted Predictors from Table 3.1

The list of predictors removed from models 2-5:

- *rsuper, persinc2, work, knowtg, polpart2, knowgl, hincpast, famrelig, tunionsa, tea, healthyr, eqnow7, eqnow9, rmarstat, umineth, ssexmarr, livearea, carehome*

7.4 Model Assessments Techniques

- Significance level of individual predictors:
  `anova(model_1, test="Chisq")`
- Multicollinearity:
  `vif(lm(outcome ~ predictors, data=data))`
- Deviance between models:
  `anova(model_1, model 2, test="Chisq")`
- Code from the lecture notes is used to generate confusion matrices & perform outlier analysis.

7.5 Handling Missing Data

In the complete case analysis, all rows with missing values were dropped.

In the other cases:

- <4%: removed from analysis as minor missing data.
- ≥4%: missing level treated as level of its own (missing) or merged with other levels where the sample size is too low. Merged levels are included in the following section.

7.6 Merged Levels

Predictor	Level	Merged Level	Notes
household	1	1	
	2	2	
	3	>2	
	4		
	5		
	6		
	7		
	8		
rmarstat	Single	Single	
	Married	Married	
	Living as married	Living_as_married	
	Separated	Separated_or_divorced	Similar %, confirmed by basic logistic regression
	Divorced		
	Widowed	Widowed	
tenshort	Own it outright	Own_house	
	Buying with help of a mortgage or loan	Buying_with_mortgage_inc_part_rent	
	Pay part rent and part mortgage (Co ownership)		
	Rent – Housing Executive	Rent_with_exec_or_asoc	
	Rent – Housing Association		
	Rent – private landlord	Rent_from_private	
	Other	Buying_with_mortgage_inc_part_rent	1 observation only, merge
highqual	Degree level or higher	Alevel_and_higher_ed	
	Higher education		
	GCE A level or equiv		
	GCSE A-C or equiv	Good_GCSE_or_other	
	Other, level unknown		
	Unclassified		Similar %, confirmed by basic logistic regression
	GCSE D-G or equiv		
	No qualifications	Bad_GCSE_or_no_qual	
tea	15 or under	15_or_under	
	16	16_and_over	
	17		
	18		
	19 or over		
	Still at school	Still_in_school	
	Still at university	Still_in_uni	
rsect	Public sector	Public	
	Private sector	Private	

	Voluntary/charity sector	Voluntary_or_other	Similar %, confirmed by basic logistic regression
	Other		
	NA	Missing	
ansseca	Large employers and higher managerial occupations	Higher_prof	
	Higher professional occupations		
	Lower managerial and professional occupations		
	Small employers and own account workers	Self_employed	
	Intermediate occupations	Intermediate_or_low_prof	
	Lower supervisory and technical occupations		
	Semi-routine occupations		
	Routine occupations		
	Never worked and long-term unemployed	Never_worked_or_long_term_unemployed	
	Missing	N/A	
chattnd2	Several times a week	Multiple_times_a_month	
	Once a week		
	2 or 3 times a month		
	Once a month	Several_times_a_year	
	Several times a year		
	Less frequently		
	Once a year	Once_a_year	
	Never	Never	
	NA	Not_religious	Based on data dictionary
polpart2	Democratic Unionist Party (DUP)	Right_or_center	
	Ulster Unionist Party (UUP)		
	Sinn Fein	Left_or_center	
	Social Democratic and Labour Party (SDLP)		
	Alliance Party	Center	
	Other	Other_or_missing	Similar %, confirmed by basic logistic regression
	Missing		
ruhappy	Very happy	Happy_or_indifferent	
	Fairly happy		
	Can't choose		
	Fairly unhappy	Unhappy	
	Very unhappy		
healthyr	Excellent	Good	
	Good		
	Fair	Fair_or_cant_choose	
	Can't choose		
	Poor	Bad	

	Very poor		
uprejgay	Very prejudiced	Very	
	A little prejudiced	A_little	
	Other		
	Not prejudiced at all	Not	
glsocdist	0	Less_than_3	
	1		
	2		
	3	3_to_5	
	4		
	5	More_than_5	
	6		
	7		
	8		
	9		
	10		
	11		
glvis	0	0_scenarios	
	1	More_than_1_scenario	
	2		
	3		
	NA	Missing	
knowgl	Know	Know	
	Don't know	Dont_know	
	Missing	Not_sure_or_no_answer	
knowtg	Know	Know	
	Don't know	Dont_know	
	Missing	Not_sure_or_no_answer	