Individual Logistic Regression Report

Course: ST211 Applied Regression
Candidate Number: 21588
Word Count: 1629

Data: Northern Ireland Life and Times Survey, 2012
Research Question: What factors affect participants to respond “Yes” when asked whether same-sex marriages have the same rights as traditional marriages?

Contents:

Introduction

The aim of this logistic regression is to understand what factors significantly affect the binary outcome variable *ssexmarr*, "Should same-sex marriages have the same rights as traditional marriages?". The data comes from the Northern Ireland Life and Times Survey, 2012: Lesbian, Gay, Bisexual and Transgender Issues Teaching Dataset. The analysis was mainly done on demographic and attitude related variables. The result of the analysis showed that the most important predictors appeared to be the political party supported (*polpart2*), level of prejudice against gay men (*uprejgay*), and age (*rage*). See *Appendix 4.1* for all significant variables in the final model.

Exploratory Analysis

Plots

I first plotted some variables to explore the data, identify possible effects on the output variable and check dependencies between predictors.

The table below shows the projected significance levels for continuous variables after observing plots. I gave a value out of 5, taking 5 if it appears to have a very significant effect on the outcome variable *ssexmarr*, and taking 0 if it appears to have no effect.

Variable Name	Variable Type	Potential Significance (?/5)
rage	continuous	5
livearea	continuous	4
persinc2	continuous	2
glsocdist	continuous	4
househld	continuous	4
glvis	continuous – converted to categorical later	1

The cross-tabulation matrix between *uprejgay* and *glchild* showed that for lower levels in *uprejgay* there were higher within group percentages of the higher levels for *glchild*. This confirmed my assumption of dependency since the lower levels in *uprejgay* show high prejudice and high levels in *glchild* show being very uncomfortable with having a gay/lesbian child, which suggest similar meanings.

The projected significance levels were not used while building the model since they might be wrong because of differing sample sizes. Instead, I started with a model including all variables and used this information to understand the data and start thinking about data manipulation.

See *Appendix 5.1* for plotting methods.

Missing Values
For all variables with less than 10% missing values, I removed missing rows, reducing the dataset from 1064 rows to 862. Then, I turned the NAs left in categorical variables to a "missing" level. After this, the only variable left with NAs was the continuous variable *persinc2* with 21.2% missing. Since this predictor was non-significant, this column was removed from the dataset to retain some reduced rows, increasing total rows to 984. In the final model, I tried turning "missing" level back to NAs, but this model was worse, hence was not used.

Merging Levels of Categorical Predictors
Some categorical variables' levels were merged to increase significance and provide easier interpretability. The sample size, plots, and the difference in significance were considered while choosing which levels to merge. See *Appendix 3*.

Multicollinearity
To prevent multicollinearity, we checked if any variables are aliased or have high VIF values. The ones with over 10 VIF were removed. See *Appendix 1.4*.

Relevelling
Each categorical variables' baseline was relevelled to the most common level and some of them were modified later to provide better interpretation. See *Appendix 1.5*.

From Initial to Final Model

Firstly, I loaded the dataset "individual_data.csv" and observed the data. Then, I replaced all the negative values with NAs. I identified categorical columns with less than 10% missing and removed missing values – *Appendix 1.6*. After that, I applied factor() to all the categorical predictors and changed their NAs to a "missing" level. Next, I changed the level 2 to 0 in the outcome variable, making it binary. I then plotted and observed some variables. I relevelled the baseline of some categorical variables to make the coefficients more interpretable. I looked at the APC of some predictors and centred a continuous variable. I tried converting some continuous variables to categorical and ran the initial model with all variables but *persinc2* and *glvis* – *Appendix 1.2*. Then, I merged some categorical variables' levels and dealt with multicollinearity. I created a third model by additionally removing the non-significant continuous and binary variables with backward elimination using summary(). In the fourth model the non-significant categorical variables were removed using the anova() test. I tried adding back the removed aliased and non-significant variables to see if it will increase significance. Only changes that improved the model were the removal of *healthyr* and *rsect*. Then, I checked if there were any significant interactions, but there was none. I turned the "missing" level back to NAs and compared the two models, choosing the one with the "missing" level. Finally, I did predictions using my final model with a confusion matrix. See *Appendix 2* for the reasons for removing variables and *Appendix 6* for the comparisons of successive models' diagnostics. The table below details changes in each model.

Model	Predictors
0	all variables
1	*persinc2*, glvis removed (alias, missing and non-significant)
2	*work*, *rsuper* removed (high VIF values)
3	*persinc2* column removed from dataset, non-significant continuous and binary variables removed from model
4	non-significant categorical removed using anova() test
Final	more non-significant variables removed after running anova() test again
Interaction	added interaction between polpart2 and uprejgay, was not significant by anova() test compared to the Final model
NAs	changed "missing" level back to NAs

Results

The odds ratio is calculated by exponentiating the coefficient of the variable. This value shows the multiplicative effect of the predictor level on the probability of the outcome binary variable. The most significant level is *polpart24*, voting for the Social Democratic and Labor Party. This attribute increases the odds of saying "Yes" to the question of *ssexmarr* by a multiple of 3.93, one of the highest odd ratios within the significant variables. Another significant level is *glchild3*, which is answering "Neither comfortable or uncomfortable" to the question "How would you feel if your children was gay/lesbian?". Interestingly, giving this answer significantly decreases the odds by a multiple of 0.35, which is one of the lowest odds. Another interesting one is the *ruhappy8* level, which also has a high odd ratio of around 4.4, but is not significant. This value means answering "Can't choose" to the question "How happy are you?". The non-significance intuitively makes sense since there could be high variability as this answer can have different meanings for different individuals or there could be insufficient amount of data.

Furthermore, I performed Average Predictive Comparisons (APC) to some important predictors, since a non-linear model's coefficients doesn't have a straightforward interpretation. APC gives a more interpretable value that is like a coefficient. For example, the difference of the probability of *ssexmarr* taking value 1 decreases by about 72% from someone at age 18 to another person at age 96.

Comments about Data and Analysis

Values can be missing in a randomly or systematic way. The variables with the highest missing values were *persinc2*, *rsuper,* and *rsect.* As all these predictors are related to income or career, it is likely to be cause by a systematic error. For example, some people may feel uncomfortable to tell their income levels, so eliminating rows with NAs may result in a biased data. For that reason, I converted *persinc2* into a categorical variable, with a "missing" level instead of removing NAs. Similarly, *chattnd2*, the frequency of attending religious services is only asked to religious people, which means that missing values directly indicate being non-religious. Overall, I tried to retain as much data as possible to keep it representative; however, there could be additional bias in data collection methods which could potentially affect the analysis and results.

A counter-intuitive result of the analysis was the variable *orient* being non-significant. People's sexual orientation should intuitively have a significant impact on whether they think same sex marriages should have the same rights as traditional marriages. This result may be caused by the small sample size of non-heterosexual participants. Furthermore, the odd ratio of *eqnow82*, 0.38 shows a decrease in *ssexmarr* probability if participant believes that women are treated fairly. Intuitively, an increase would be expected since someone thinking that women are treated unfairly might think that same sex marriages should have the same rights as traditional ones.

The final model contained 11 variables, corresponding to 28 predictors including categorical levels. The difference between the residual and null deviance was 493.1, which is higher than the first few models' ones, even though they have over 75 predictors. The removal of *persinc2* significantly improved this diagnostic. The Confusion Matrix for the final model showed an 81% accuracy of the predictions. See *Appendix 4*.

Lay Report

"DO YOU THINK SAME SEX MARRIAGES SHOULD HAVE THE SAME RIGHTS AS TRADITIONAL MARRIAGES?"

Recent research done on the factors affecting people's opinion about the rights of same sex marriages produced some interesting results. The researchers tried to understand what demographic features or personal attributes determine whether someone thinks that same sex marriages should have the same rights as traditional marriages. About 1000 people were surveyed in Northern Ireland in 2012 and then the results were analysed to create a statistical model which can estimate people's likelihoods of answering "Yes" or "No" to the question "Do you think same sex marriages should have the same rights as traditional marriages?". The results showed that the most important factors were their age, the political party they support, and whether they are prejudiced against gay men. Other additional important factors include how happy they are currently, their genders, how often they attend religious services, and their opinions on various questions around LGBTQ topics.

Amy

Amy is a woman at age 25, with a Bachelor's Degree, who believes that women are generally treated unfairly. She attends religious services several times in a year and supports the Social Democratic and Labor Party. In the survey she said that she is feeling very happy these days. She believes that lesbians and gay man are born that way and mentions that she would be very comfortable with having a gay or lesbian child. Moreover, she never experienced a situation where she felt uncomfortable if someone around her was gay or lesbian and says that she is not prejudiced against gay man at all.

Davis

He is a man at age 60, with no educational qualification, who does not believe that women are generally treated unfairly. He attends religious services once a week and supports Democratic Unionist Party. He mentioned that he is not very happy these days. He believes that lesbians and gay man choose to be that way and mentions that he would be very uncomfortable with having a gay or lesbian child. Furthermore, he experienced 5 situations where she felt uncomfortable around gay or lesbian people and says that he is very prejudiced against gay man.

Probability of saying "yes":

0.99 **0.013**

Appendix

Appendix 1: Chronological Step of Building the Model

1.1 Model Building Steps Table

	Methodology
1	Loading and observing the data
2	Changing negative values with NA
3	Removing all missing values from categorical predictors with less than 10% missing values
4	Converting all categorical variables to factor
5	Changing categorical variables' NAs to a "missing" level
6	Change level 2 to 0 in *ssexmarr*
7	Plotting continuous variables
8	Plotting categorical variables
9	Relevelling the baseline for categorical variables
10	Looking at APC of some important predictors
11	Centering some continuous predictors
12	Converting continuous *glvis* to categorical *factor_glvis*
13	Running initial model with all variables but *persinc2* and *glvis*
14	Merging levels in categorical variables to make it more significant
15	Dealing with multicollinearity (*rsuper* and *work* removed)
16	Removing *persinc2* column from the dataset (not used in model)
17	Removing non-significant continuous and binary variables using summary()
18	Removing non-significant categorical variables using anova() test
19	Tried adding back removed variables and removing current variables, compared using anova() and made a few changes
20	Checking interactions
21	Replacing all "missing level" with NA and compare models
22	Predicting outcomes using the final model and checking confusion matrix

1.2 Step 12&13 Explanation

The variable *glvis,* which shows the number of scenarios where same sex couples would be more visible, was initially a non-significant continuous variable. I turned it into a categorical variable *factor_glvis.* This data manipulation increased the difference between residual and null deviance in the initial model and increased the significance. I also tried merging levels of *factor_glvis* but it did not make the model better, hence was not done at the end. Even though this variable appeared to be more significant than the initial continuous one, it was not used in the final model as it showed to be non-significant at the end. Moreover, *persinc2* was a continuous predictor with lots of missing values. This was also converted to a categorical variable *factor_persinc2*, but it was still not significant and instead of doing this, removing the *persinc2* column from the dataset gave much better model diagnostics in the later steps.

1.3 Step 20: Interactions
I tried interactions between the top three most significant variables in the final model, which were *glchild, polpart2,* and *uprejgay.* I first ran separate regressions with only the interaction predictors. Only the interaction between polpart2 and uprejgay was significant. I tried adding this interaction to the final model, but since it was not significant in that model, I did not end up including it.

1.4 Step 15: Dealing with Multicollinearity
Initially, *ansseca* was an aliased variable. However, after the data manipulations, it was not alised anymore. In model 0, *factor_glvis* was aliased with *glvis* as expected since the model contained both the original and the modified variable. After *glvis* was removed, there were a few left with high VIFs, so first *rsuper* was removed and then *work* was removed. Interestingly *rsect* initially had a high VIF value, but after the removal *rsuper* and *work*, the VIF value decrease to about 1. This shows that it might have been collinear with one of the removed variables.

1.5 Step 9: Relevelling the baseline of Categorical Variables
Initially, all categorical variables' baselines were set to the most common level. However, to better interpret some levels' significance levels and coefficients, further relevelling was done to certain variables. For example, in *religcat* and *famrelig*, the baseline was initially "Catholic", so I turned it to "No religion", so that the difference between a religious and non-religious person can be seen instead of the difference between a Catholic and Protestant. Similarly, *rage*, which shows the age of the participant, had a y-intercept 0, which realistically did not make sense. Therefore, it was centred to the average age. However, doing APC on this predictor afterwards provided even better interpretability.

1.6 Step 3: Removing NAs from Categorical Variables with less than 10% Missing
I initially removed the NA rows from all categorical predictors with less than 10% missing data. However, after reaching the final model, I went back and excluded all non-significant categorical predictors from this removal to retain as much data as possible.

1.7 When did you decide to remove a predictor?
We first decided to remove *persinc2* while dealing with missing values since it had 19% NAs and was non-significant. After deciding to convert *glvis* into a categorical variable *factor_glvis* we removed *glvis* to prevent multicollinearity. Then we removed some variables because of high VIF values. Next I removed non-significant continuous and binary variables with a backward elimination and then did the same for categorical variables. Finally, I removed each variable one by one from the model and compared with the initial model using anova() test. I ended up removing two more variables from the final model.

Appendix 2: Removed Variables and Reasons

Removed Variable Name	Reasoning for Removal
persinc2	19% missing values and non-significant
glvis	Turned to categorical variable *factor_glvis*
rsuper	High VIF (188)
work	High VIF (33)
umineth	Non-significant continuous/binary variable from summary()
eqnow9	Non-significant continuous/binary variable from summary()
livearea	Non-significant continuous/binary variable from summary()
anyhcond	Non-significant continuous/binary variable from summary()
knowgl	Non-significant continuous/binary variable from summary()
intwww	Non-significant continuous/binary variable from summary()
religcat	Non-significant continuous/binary variable from summary()
eqnow3	Non-significant continuous/binary variable from summary()
tunionsa	Non-significant continuous/binary variable from summary()
househld	Non-significant continuous/binary variable from summary()
carehome	Non-significant categorical using anova()
factor_glvis	Non-significant categorical using anova()
knowtg	Non-significant categorical using anova()
ansseca	Non-significant categorical using anova()
hincpast	Non-significant categorical using anova()
rmarstat	Non-significant categorical using anova()
tenshort	Non-significant categorical using anova()
eqnow11	Non-significant categorical using anova()
tea	Non-significant categorical using anova()
healthyr	Non-significant in final model using anova()
rsect	Non-significant in final model using anova()

Appendix 3: Data Manipulation

3.1 Merging Method

The levels with smaller sizes were mostly merged with other levels. I tried not to merge two levels with big sample sizes. The boxplots were checked to see which levels illustrate similar patterns or values for the outcome variable. After each merging, the change in the difference

between residual and null deviances were compared to the critical value for chi-square distribution with degrees of freedom that is the difference in the number of predictors in the model. If the decrease in deviance difference was bigger than the critical value, then this showed that it did not become more significant, hence the merging was undone. See Appendix 3 for more detail.

3.2 Merged Levels Table

Variable name	Merged Levels	Included in the model?	Significance before/after merging
rmarstat	2&3	No	not significant at 5% level but increased significance
tenshort	6&7	No	already significant at 5% level, increased significance
highqual	1&2, 4&5, 6&7	Yes	not significant at 5% level but increased significance
tea	1&2&3&4, 6&7&8	No	not significant at 5% level but increased significance
religcat	1&2	No	not significant at 5% level but increased significance
famrelig	1&2	No	was not significant at 5% level, merging made it significant
chattnd2	1&2, 3&4, 5&6&7	Yes	not significant at 5% level but increased significance
orient	2&3&4	No	not significant at 5% level but increased significance
polpart2	1&3, 2&5&6&7	Yes	was not significant at 5% level, merging made it significant
healthyr	2&3&4&5	No	was not significant at 5% level, merging made it significant
glchild	1&2	Yes	was not significant at 5% level, merging made it significant
work	3&4	No	already significant at 5% level, increased significance
rsect	3&4	No	not significant at 5% level but increased significance

3.3 Merged Levels Extra Notes

famrelig & *religcat*: Merged levels catholic and protestant, so turned into a binary variable with levels religious and not religious

orient: Levels 3 and 4 have very few values, so merged with level 2 and turned into a binary variable with levels heterosexual or not heterosexual

polpart2: The only significant level is 4, so reducing the number of levels by merging some levels made the variable more significant

Appendix 4: Final Model Variables & Significance

4.1 Final Model Variables

The final model includes 11 significant variables: whether they think homosexuality comes from birth or is chosen (*glborn*), number of situations would feel uncomfortable if person was gay/lesbian (*glsocdist*), opinions on having a gay/lesbian child (*glchild*), their prejudice against gay men (*uprejgay*), happiness level (*ruhappy*), supported political party (*polpart2*), frequency of attending religious services (*chattnd2*), whether they think women are treated unfairly (*eqnow8*), highest educational qualification (*highqual*), gender (*rsex*), and age (*rage*).

4.2 Table of Final Model Variables/Levels

Variable Name/Level	Odd Ratio	Pr(>|z|)	Significance
glborn2	0.45153898	0.000990	***
glbornmissing	0.54557831	0.041536	*
glsocdist	0.85015188	0.005765	**
glchild3	0.35219004	1.96e-05	***
glchild4	0.28772040	0.000938	***
glchild5	0.30281454	0.015342	*
uprejgay2	0.83986643	0.724998	
uprejgay3	2.49828736	0.054776	.
uprejgay4	1.82747723	0.585215	
ruhappy2	0.80182212	0.275329	
ruhappy3	0.51161588	0.056447	.
ruhappy4	0.05328915	0.001283	**
ruhappy8	4.41025090	0.254402	
polpart24	3.93427503	2.24e-05	***
polpart26	1.89421008	0.003139	**
polpart2missing	2.18647882	0.021465	*
chattnd23	2.28186513	0.009513	**
chattnd27	1.74987825	0.022357	*
chattnd28	3.37090347	0.000146	***
chattnd2missing	2.17161556	0.008381	**
eqnow82	0.38998327	0.039824	*
highqual3	1.12062899	0.682880	
highqual4	0.69383594	0.159663	
highqual6	0.83695289	0.492819	
highqual9	3.33354724	0.041259	*
rsex2	1.60843045	0.011900	*
rage	0.96448394	1.25e-09	***

4.3 Final Model Confusion Matrix Diagnostics

% Correct (0): 0.67
% Correct (1): 0.9
% Total Accuracy: 0.81

Appendix 5: Plotting

5.1 Plotting Methods

The continuous variables were plotted by creating bins, calculating the corresponding proportion of the binary output variable taking value 1, creating a new data frame containing the midpoint of each bin, and fitting a generalised linear model which illustrates a logistic regression curve. For categorical variables, boxplots were created with the output variable on the x-axis, showing the range of probabilities and the most common output value for each level.

5.2 Cross-Tabulation Plot between *uprejgay* and *glchild*

The variable *uprejgay* shows the participants' level of prejudice against gay men and the variable *glchild* shows how comfortable the participant will be if their child was gay/lesbian. Intuitively, these two variables can have a potential dependency on each other, as they have similar meanings.

Appendix 6: Model Diagnostics Comparison

Model No	**0**	**1**	**2**	**3**	**4**	**Final**	**Interaction**	**NAs**
Residual Deviance	529.3	531.1	546.4	737.6	775.8	789.1	774.4	526.6
Null Deviance	1006.9	1006.9	1006.9	1282.2	1282.2	1282.2	1282.2	857.8
Deviance Difference	477.6	475.8	460.5	544.5	506.4	493.1	507.8	331.2
# of variables	85	84	79	66	35	28	36	25