University of Oxford: MPhil in Politics

1090063

2025-01-16

Contents

1 Exercise 1: Owning a gun in the U.S.

1.1 Data Summary

The Cooperative Congressional Election Study surveys Americans during elections. The data collected from 2022 provides useful insights into gun ownership in the U.S. The following analyses is based on the `cces22.csv` file consisting of 60,000 observations over 9 variables.

The variables used in the analysis are coded as: `age`, `female`, `register`, `white`, `pid`, `trust_fed_gov`, `region`, and `education`. Gun ownership is a binary variable where `Gun Ownership = 1` and `No Gun = 0` and has been converted to a factor variable for readability. The average age of respondents is 50.4. This exercise is predominantly interested in gun ownership. By summarising the data as percentages, we can understand the distribution of gun ownership across different groups. *Figure (1)* shows how gun ownership is distributed by political affiliation.

The next section looks at building an appropriate model for predicting gun ownership and producing a clear demographic profile of gun owners in the U.S.

1.2 Model of Gun Ownership

To model gun ownership, we initially take into account the following variables: `age`, `female`, `white`, `region`, and `education`. Consequently, the model is specified in (1):

$$Y_{\text{Gun=1}} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \beta_3 \cdot \text{female} + \beta_4 \cdot \text{white} + \beta_5 \cdot \text{region} + \beta_6 \cdot \text{education} + \epsilon \quad (1)$$

As age is a continuous variable, it cannot be assumed that there is a linear relationship between `age` and $Y_{\text{Gun=1}}$. Therefore, a quadratic term age^2 is included to account for non-linear relationships. This non-linear assumption is supported by Figure (2) showing the percentage of respondents owning guns by age to be slightly concave in nature.

To estimate this model, logit or probit models are most appropriate due to the binary, categorical nature of the dependent variable,$Y_{\text{Gun=1}}$. Therefore, (assuming a logit approach) the model is as follows:

$$\text{logit}(P(\text{gun} = 1)) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \beta_3 \cdot \text{female} + \beta_4 \cdot \text{white} + \beta_5 \cdot \text{region} + \beta_6 \cdot \text{education} + \epsilon \quad (2)$$

where `logit` is the logistic function, `P(gun = 1)` is the probability of gun ownership, and ϵ is the error term.

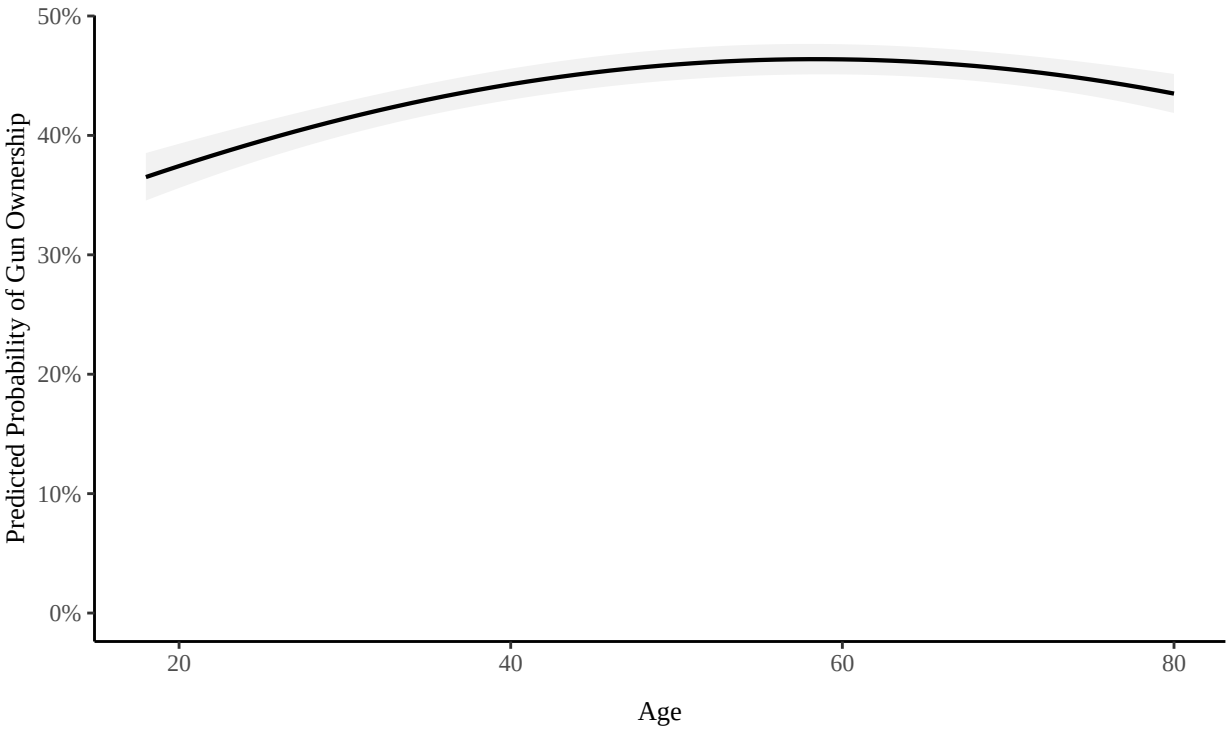
Initially, a few incomplete models were run to check for the significance of individual variables. The results showed that `age` and `white` were significant predictors of gun ownership. For example, a logit model based just on `white` showed the predicted probability of gun ownership is higher for white individuals (41%) compared to non-white individuals (30%).

Next, the full model containing `age`, `I(age^2)`, `female`, `white`, `region` and `education` was estimated for both logit and probit model to check for consistency in direction of effects. *Table (1)* shows these results,

comparing both logit and probit models. The primary focus is on the logit regression results which are given as log-odds in the first column. The odds ratio for `age` is `1.03`, indicating that for each additional year of age, the odds of gun ownership increase by 3%. The odds ratio for `white` is `1.68`, indicating that white individuals have `68%` higher odds of owning a gun compared to non-white individuals. Therefore, with taking the modal values of categorical variables into account, the model predicts the probability of gun ownership based on the demographic profile of the average respondent. This demographic profile indicates that white men in the South with a 2-year college degree are most likely to own guns, compared to baseline alternatives of being a woman, non-white, living in the South, and being a high school graduate.

To focus on age more specifically, the predicted probabilities of gun ownership by age are shown in *Figure (3)*. The plot shows that the predicted probability of gun ownership increases with age up to around 57 years before stabalising and declining slightly again. The shaded region represents 95% confidence intervals around the predicted probabilities. The predicticted probability of owning a gun peaks at around age 57 when estimated as a function of `age`, `I(age^2)`, `female`, `white`, `region` and `education`.

Figure 1: Predicted Probability of Gun Ownership by Age



Notes: The plot shows the predicted probabilities of gun ownership by age, estimated from a logistic regression model. The shaded region represents 95% confidence intervals. Predictors with p−values < 0.05 are considered statistically significant. Predictions are adjusted for gender, race, region, and education.

1.3 Considering Political Variables

Next, the model is re-estimated to consider variables related to politics like partisanship, trust in government, and whether the respondent was registered to vote (`pid`, `trust_fed_gov`, `register`).

An updated table comparing the logit and probit models is provided in *Table (2)* below.

Table 1: Comparison of Logit and Probit Models with Log-Odds

	Logit (Odds Ratio)	Probit (Latent z-scores)	Logit (Log-Odds)
Intercept	0.312***	−0.711***	−1.166***
	(0.029)	(0.056)	(0.093)
Age	1.030***	0.018***	0.029***
	(0.004)	(0.002)	(0.004)
Age Squared	1.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)
Female (1 = Female)	0.702***	−0.218***	−0.354***
	(0.014)	(0.012)	(0.020)
White (1 = White)	1.684***	0.319***	0.521***
	(0.039)	(0.014)	(0.023)
Region: Northeast	0.415***	−0.537***	−0.880***
	(0.012)	(0.017)	(0.029)
Region: Midwest	0.755***	−0.173***	−0.281***
	(0.019)	(0.016)	(0.025)
Region: West	0.715***	−0.206***	−0.336***
	(0.019)	(0.016)	(0.026)
Education: Some College	1.083**	0.049**	0.079**
	(0.031)	(0.017)	(0.028)
Education: 4-Year Degree	0.906***	−0.060***	−0.099***
	(0.025)	(0.017)	(0.028)
Education: 2-Year Degree	1.225***	0.126***	0.203***
	(0.042)	(0.021)	(0.034)
Education: Postgraduate Degree	0.810***	−0.128***	−0.211***
	(0.026)	(0.020)	(0.032)
Education: No High School Diploma	0.763***	−0.163***	−0.270***
	(0.048)	(0.038)	(0.063)
Num.Obs.	47 695	47 695	47 695
RMSE	0.47	0.47	0.47

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Standard errors are shown in parentheses. Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. The baseline for regional comparisons is the South region. The baseline for education comparisons is 'High school graduate' as these are the modal values for each variable.

Table 2: Comparison of Logit and Probit Models with Log-Odds (inc. political variables)

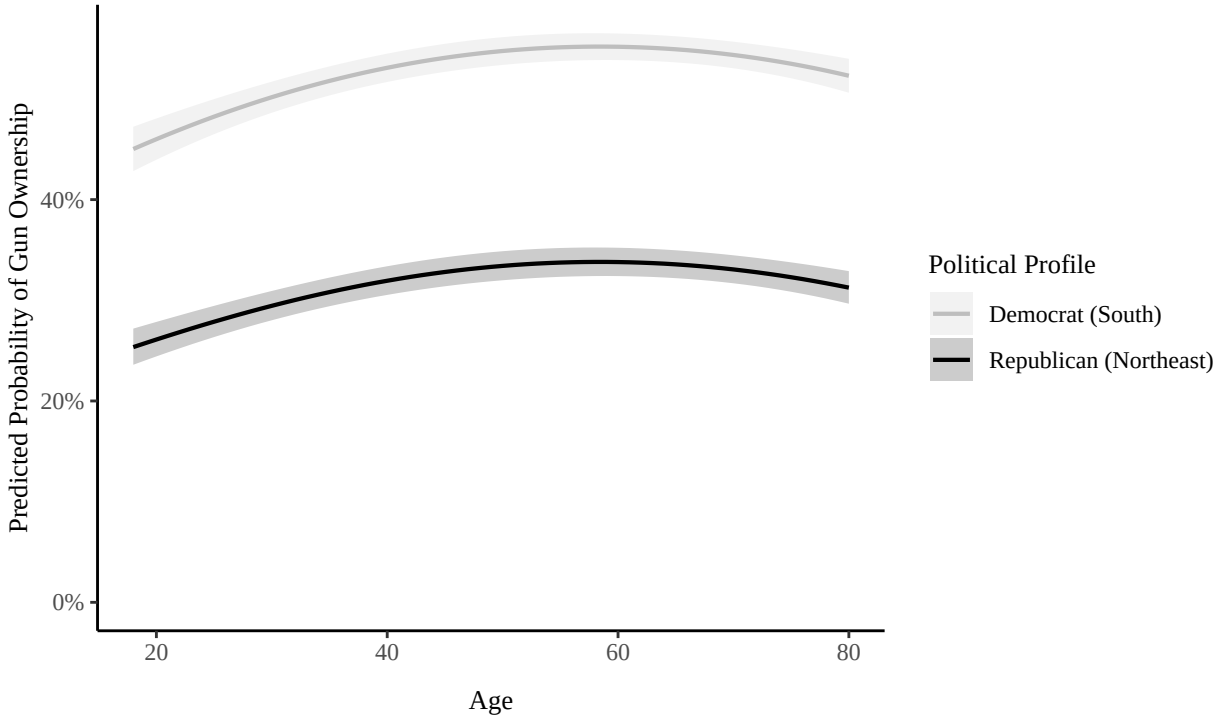
	Logit (Odds Ratio)	Probit (Latent z-scores)	Logit (Log-Odds)
Intercept	0.142***	−1.949***	−1.949***
	(0.015)	(0.104)	(0.104)
Age	1.026***	0.025***	0.025***
	(0.004)	(0.004)	(0.004)
Female (1 = Female)	0.749***	−0.288***	−0.288***
	(0.015)	(0.020)	(0.020)
White (1 = White)	1.386***	0.326***	0.326***
	(0.034)	(0.024)	(0.024)
Region: Northeast	0.433***	−0.837***	−0.837***
	(0.013)	(0.030)	(0.030)
Region: Midwest	0.786***	−0.241***	−0.241***
	(0.021)	(0.026)	(0.026)
Region: West	0.742***	−0.298***	−0.298***
	(0.020)	(0.027)	(0.027)
Education: Some College	1.112***	0.107***	0.107***
	(0.033)	(0.029)	(0.029)
Education: 2-Year Degree	1.246***	0.220***	0.220***
	(0.044)	(0.036)	(0.036)
Political Affiliation: Republican	2.460***	0.900***	0.900***
	(0.068)	(0.028)	(0.028)
Political Affiliation: Independent	1.616***	0.480***	0.480***
	(0.043)	(0.026)	(0.026)
Political Affiliation: Other	1.673***	0.514***	0.514***
	(0.082)	(0.049)	(0.049)
Trust in Government: A Great Deal	0.867**	−0.142**	−0.142**
	(0.040)	(0.046)	(0.046)
Trust in Government: A Fair Amount	0.828***	−0.189***	−0.189***
	(0.021)	(0.025)	(0.025)
Trust in Government: None at All	1.333***	0.288***	0.288***
	(0.035)	(0.027)	(0.027)
Not Registered to Vote	1.837***	0.608***	0.608***
	(0.082)	(0.044)	(0.044)
Num.Obs.	47 283	47 283	47 283
RMSE	0.46	0.46	0.46

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Standard errors are shown in parentheses. Statistically insignificant variables have been removed. The baselines for region, education, political affiliation, and registered are 'South', 'High school graduate', 'Democrat' and 'Registered' respectively.

This updated model now shows how demographic traits from the original model have been dampended by the political variables which have a large effect. For example, the odds ratio for being a Republican is `2.46`, indicating that Republicans have `146%` higher odds of owning a gun compared to Democrats. Moreover, having some level of trust in government reduces the odds of owning a gun compared to having not much trust, whereas not being registered to vote increases the odds of owning a gun by 84%, compared to being registered.

Next, we look specifically at certain profiles: a registered white male Republican from the North-East, compared to a registered white male Democrat from the South. These results are shown together in Figure (4) below. For the Republican profile, the predicted probability of gun ownership at its peak (60 years of age), is at `55%`, whereas for the Democrat profile (also at 60 years of age), the predicted probability is at `34%`. This shows how political affiliation is such a strong driver of gun ownership in the U.S.

Figure 2: Predicted Probability of Gun Ownership by Age for Two Political Profiles



Notes: The plot shows the predicted probabilities of gun ownership by age for two political profiles, estimated from a logistic regression model. The shaded regions represent 95% confidence intervals. Predictions use the modal values of education and trust in government, with the other variables specified for each profile.

2 Exercise 2: Preferences for redistribution across the EU

2.1 Data Summary

The dataset used in the following analysis comes from Round 9 (2018) of the European Social Survey (ESS). The primary dependent variable of interest is `red`, which measures preferences for redistribution on a

numerical scale from 1-5 (1 = low, 5 = high). The average age of respondents is 52.4 and average preference for redistribution is 3.76. A histogram plot of the values `red` can take is plotted below in *Figure (5)*. The dataset contains a range of demographic variables which may influence preferences for redistribution which are analysed in the following sections.

2.2 Ordered Logit Model

The first model estimated is an ordered logit model to predict preferences for redistribution based on all of the available variables. This ordered logit model is used as `red` is oridinal (values from 1-5), but we are unsure on whether the intervals are equally spaced. The model provides an estimate of the probability that the outcomes `red` falls into one of the ordered categories by modelling the log-odds of being at or below a certain category `j`. The model is specified as follows:

$$\text{red} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{mbtru} + \beta_3 \cdot \text{imbgeco} + \beta_4 \cdot \text{ctzcntr} + \beta_5 \cdot \text{gndr} + \beta_6 \cdot \text{iphlppl} + \beta_7 \cdot \text{hhmmb} + \beta_8 \cdot \text{rlgblg} + \epsilon \quad (3)$$

The results of the ordered logit model are shown in *Table (3)* below. The table shows the ordered logit model estimates for how different variables in the model influence respondents' preferences for income redistribution, measured on a scale from 1 (low support) to 5 (high support).

Firstly, age, citizenship status, and the number of household members have no statistically significant effect on preferences for redistribution. However, being a member of a trade union (both previously and currently) has a significant positive effect on preferences for redistribution. Those who see immigration as beneficial to the economy are more likely to support redistribution. Whereas men are less likely to support redistribution than woman, as are those with religious beliefs. Finally, altruism and income have a positive effect on preferences for redistribution.

2.3 Expected Probabilities

Although the effect of opinions on the impact of immigration on the economy `imbgeco` was previously left as a continuous variable, we now consider the expected probabilities of preferences for redistribution based on this variable from an ordinal perspective. The plot below in *Figure (6)* shows the predicted probabilities of preferences for redistribution across each level of `imbgeco` from 0 to 10.

Whilst there my be hypotheses that those who think immigration is beneficial for the economy will want more redistribution to encourage immigration, the results show that support for redistribution is generally consistent across different levels of perceived impact of immigration on the economy. The greatest increase in support for redistribution as `imbegco` increases is seen for those with the highest level of support for redistribution. The results suggest that the majority of the preferences towards redistribution are moderate, suggesting how other factors than just attitudes on immigration effect redistribution attitudes. The results are all statistically significant, with narrow confidence intervals.

Table 3: Ordered Logit Model for Preferences for Redistribution

	Ordered Logit Model
Low to Moderate-Low Cut	−4.324***
	(0.324)
Moderate-Low to Moderate Cut	−2.031***
	(0.271)
Moderate to Moderate-High Cut	−0.851**
	(0.265)
Moderate-High to High Cut	1.152***
	(0.267)
Age	−0.005
	(0.003)
Member of Trade Union (Previously)	0.252*
	(0.108)
Member of Trade Union (Currently)	0.482***
	(0.131)
Effect of Immigration on Economy	0.052**
	(0.019)
Not a Citizen	0.127
	(0.201)
Male	−0.357***
	(0.091)
Altruism	0.101***
	(0.023)
Household Members	0.061
	(0.041)
Religious Beliefs = Yes	−0.290**
	(0.093)
Income	−0.099***
	(0.017)
Num.Obs.	1765
RMSE	3.70

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Standard errors are shown in parentheses. Note that 'imbgeco' has been left as a continuous variable rather than being converted into an ordinal variable. Whilst it has ordered properties, it is assumed linear here.

Table 4: Brant Test for Proportional Odds Assumption

Predictor	Chi-Square	Degrees of Freedom	p-value
Omnibus	54.87	30	0.0037
Age	0.10	3	0.9914
Union Membership (Previously)	1.38	3	0.7091
Union Membership (Currently)	5.94	3	0.1145
Perception of Immigration Impact	2.41	3	0.4919
Non-Citizen	0.38	3	0.9450
Gender (Male)	6.08	3	0.1077
Altruism	12.73	3	0.0053
Household Size	6.60	3	0.0857
Religious Belief (Yes)	10.79	3	0.0129
Income Level	10.82	3	0.0127

Note:
The null hypothesis assumes the proportional odds assumption holds. A p-value < 0.05 suggests violation of the assumption.

2.4 Proportional Odds Assumption

When we have an ordinal dependent variable, we can model the outcome using an ordered logit model. We need a number of key assumptions to hold for this model to estimate unbiased, valid results. The primary assumption of this model is the proportional odds assumption. This assumption states that each effect of the indepdendent, predictor, variables is constant across all levels of the dependent variable. In other words, the slope coefficent is the same for all categories of the dependent variable, meaning we only need to predict one set of β coefficients for all levels of the dependent variable, helping interpretability. However, if this assumption does not hold, we will have biased estimates as the effect of predictors differs across outcome levels, such that the model misrepresents the data.

To test this assumption, we use the `brant` function from the `brant` package in R. The results of the Brant test are shown in *Table (4)* below.

To interpret the `brant` test, we look at the p-values for each variable. If the p-value is $p \leq 0.05$, we reject the null hypothesis that the proportional odds assumption holds, and the assumption is violated. The `Omnibus` value gives an overall test of the proportional odds assumption for the whole model.

The variables for altruism `(iphlppl)`, religious beliefs `(rlgblg)`, and income `(inc_c)` all have p-values less than $p \leq 0.05$. This suggests that the effect of these predictors on redistribution preferences varies across categories. The overall test of the proportional odds assumption, given by `Omnibus` is also violated.

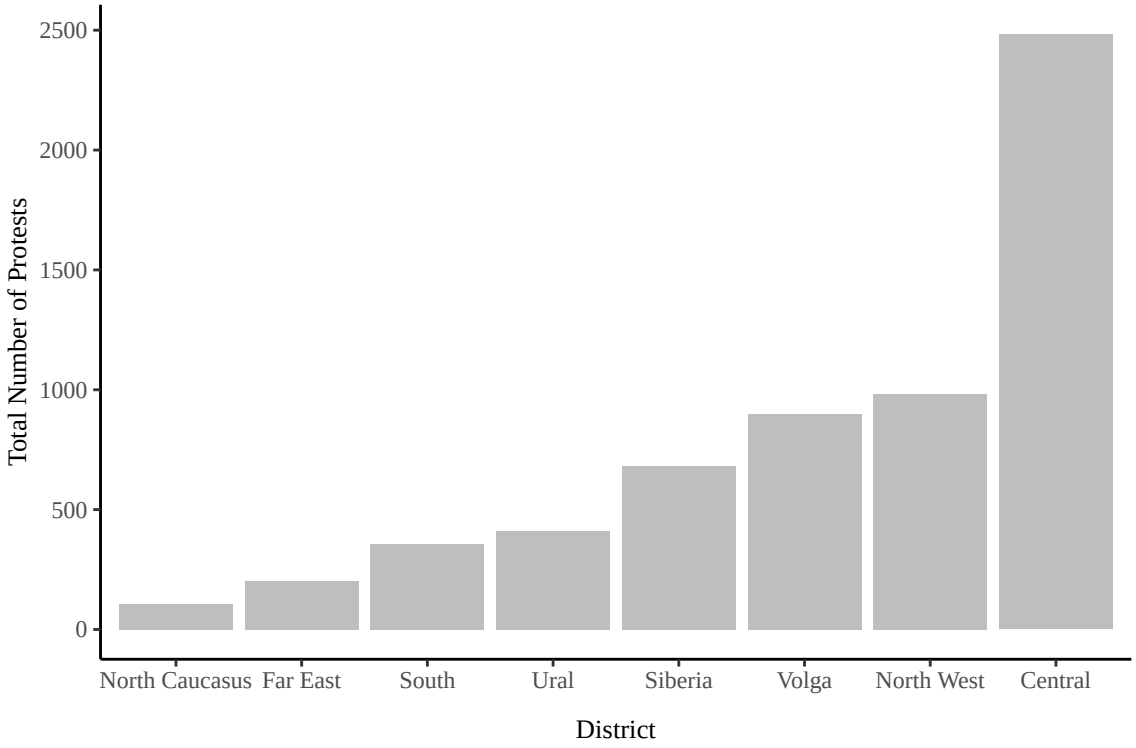
From the `ordered_logit_model` used throughout these estimates, the parallel lines assumption is violated overall and the use of an ordered logit model may not be entirely appropriate, with a generalised ordered logistic model being a possible alternative.

3 Exercise 3: International trade and protests in Russia

3.1 Data Summary

The dataset used in the following analysis, `protests.dta`, contains ~2,000 observations of protests in Russia. The data contains variables on the Russian regions and districts where protests occurred, along with socio-economic and demographic variables. The primary dependent variable of interest is `protest_ikd` which gives the number of protests in a given region for a specific year. The key predictor of interest is `tradeopen`, which measures the sum of imports and exports (% of GDP). For example we can see the count of the number of protests by district in *Figure (3)* below.

Figure 3: Total Number of Protests by District



3.2 Understanding Count Models

An OLS regression model is typically used to model the relationship between a continuous dependent variable and one or more independent variables. For OLS to be appropriately used, the residuals should be normally distributed and homoscedastic in nature. However, when the dependent variable is not continuous, other models should be used. Specifically, in the case of count data, where the dependent variable is a non-negative integer, count models are more appropriate. Count data refers to data that measures how often something (e.g., protests, `protest_ikd`) occurs in a given time interval or space, measuring the frequency of the event. Therefore, count models such as Poisson or Negative Binomial models can be used to model the relationship between the count dependent variable and the independent variables, instead of an OLS model.

These count models look at calculating the probability of observing a certain count given the independent variables, based on the Poisson distribution. A Poisson model specifically assumes that the dependent variable, Y follows the Poisson distribution of parameter rate (expected occurrences) λ such that $Y \sim \text{Poisson}(\lambda)$. This means that the mean and the variance of the dependent variable are equal: $E(Y) = \mu = \text{Var}(Y)$. However, in practice, the variance of count data is often greater than the mean ($\text{Var}(Y) > E(Y)$) known as overdispersion. In such cases, a Negative Binomial model is used which builds on the Poisson distribution to allow for the variance to be greater than the mean, providing a better fit for the data.

With the data on protests being count data, and skewed such that the errors will not be normally distributed, a Poisson or Negative Binomial model is more appropriate than an OLS model. ## Protests between 2000-2014

We are firstly interested in the relationship between international trade openness and the number of protests in Russia between 2000 and 2014. To determine whether a Poisson or Negative Binomial model is most appropriate, two overdispersion tests are conducted. Firstly, a test of a baseline negative binomial model using `odTest()` is used to show that the Null Hypotheses (no overdispersion) is rejected and therefore a negative binomial model is more appropriate. Moreover, by evaluating the mean and variance values across 2000-2014 for our dependent variable `protest_ikd`, we find that they are significantly different (`mean = 12.28 < var = 1507.53`), again supporting the use of a negative binomial model.

As a negative binomial model is most appropriate, this is now used to estimate the relationship between international trade openness and the number of protests in Russia between 2000 and 2014. The coefficient for `tradeopen` is not statistically significant in predicting the number of protests. This suggests that levels of trade openness do not have a significant effect on the number of protests in Russia between 2000 and 2014.

However, a number of regions have strong statistically significant effects, suggesting that protests are not uniform across Russia, but rather vary significantly by region. For example, `Leningrad`, `Kaliningrad`, and `Moscow` have a significantly higher number of protests compared to the omitted reference category of `Central` Russia. This suggests that regional factors play a more significant role in predicting the number of protests than trade openness.

3.3 Interaction Effects of Educational Attainment and Trade Openness

After initially finding that there is little explained effect of trade openness on protests, we are now interested in seeing whether the effect of trade openness on protests varies depending on a region's educational attainment. In other words, does educational attainment modify the relationship between trade openness and protest activity.

We are primarily interested in whether the variables `seceduc`, `tereduc` and `seceduc + tereduc` have a moderating effect on the relationship of the independent variable `tradeopen` and dependent variable `protest_ikd`. We therefore need to test different models with education variables interacted with `tradeopen` to see if they have a significant effect on the number of protests.

$$\text{Protest Count} = \beta_0 + \beta_1 \text{tradeopen} + \beta_2 \text{seceduc} + \beta_3 \left(\text{tradeopen} \cdot \text{seceduc}\right) + \varepsilon \tag{4}$$

Table 5: Negative Binomial Regression: Trade Openness and Significant Regions

	(1)
Trade Openness	−0.007
	(0.006)
Moscow (City)	6.533***
	(0.749)
Leningrad Region	4.849***
	(0.823)
Saint Petersburg	4.580***
	(0.712)
Kaliningrad Region	4.169***
	(0.969)
Novosibirsk Region	3.729***
	(0.586)
Sverdlovsk Region	3.643***
	(0.608)
Moscow (Oblast)	3.603***
	(0.616)
Samara Region	3.592***
	(0.610)
Irkutsk Region	3.534***
	(0.611)
Perm Region	3.513***
	(0.608)
Num.Obs.	480
RMSE	12.02

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

Standard errors are shown in parentheses. The table shows the significant regions with the largest effects on the number of protests in Russia between 2000 and 2014.

$$\text{Protest Count} = \beta_0 + \beta_1 \text{tradeopen} + \beta_2 \text{tereduc} + \beta_3 \left(\text{tradeopen} \cdot \text{tereduc}\right) + \varepsilon \tag{5}$$

$$\text{Protest Count} = \beta_0 + \beta_1 \, \text{tradeopen} + \beta_2 \left(\text{seceduc} + \text{tereduc}\right) + \beta_3 \left(\text{tradeopen} \cdot \left(\text{seceduc} + \text{tereduc}\right)\right) + \varepsilon \tag{6}$$

These models are estimated again using a negative binomial model, to account for the count nature of the dependent variable and overdispersion in the data. The results of these models are shown in *Table (6)* below. From the results, we see that in the model of the interaction between trade openness and secondary education, trade openness is positive and statistically significant (unlike without the interaction term). This means that higher trade openness can be associated with an increase in protests in the regions where there is low secondary education, with the `tradeopen:seceduc` interaction term showing that as secondary education increases, the effect of trade openness on protests decreases. This suggests that the effect of trade openness on protests is moderated by the level of secondary education in a region. For the second model with tertiary education, there is an association between a decrease in protests in regions with lower tertiary education. The final model suggests that trade openness decreases the number of protests in regions where there is a low overall level of education. In regions where the combined level of education is higher, trade openness is more likely to effect the number of protests. Consequently, we can conclude that where there are lower levels of education, trade openness is more likely to increase the number of protests.

The models specified, however, assume that the effects of trade openness are linear across all levels of education, and the effects of secondary and tertiary education are of equal weight. For example, the added value of tertiary education could have a much stronger effect in removing someone from being sheltered from the effects of trade openness. Firstly a weighted model is given by:

$$\text{Protest Count} = \beta_0 + \beta_1 \, \text{tradeopen} + \beta_2 \, \text{weightededucation} + \beta_3 \left(\text{tradeopen} \cdot \text{weightededucation}\right) + \varepsilon \tag{7}$$

where $\text{weightededucation} = \alpha \cdot \text{seceduc} + (1 - \alpha) \cdot \text{tereduc}$, with $\alpha = 0.37$, giving 70% weighting to tertiary education and 30% to secondary education. A second additional specification looks at the non-linear effects of both secondary and tertiary education to identify whether these education levels have different relationships with trade openness and protests.

$$\begin{aligned} \text{Protest Count} = \beta_0 + \beta_1 \, \text{tradeopen} + \beta_2 \, \text{seceduc} + \beta_3 \, \text{seceduc}^2 + \beta_4 \, \text{tereduc} + \beta_5 \, \text{tereduc}^2 \\ + \beta_6 \left(\text{tradeopen} \cdot \text{seceduc}\right) + \beta_7 \left(\text{tradeopen} \cdot \text{tereduc}\right) + \varepsilon \end{aligned} \tag{8}$$

From the results, we find that the weighted model provides value insight whereas the non-linear model is not statistically significant. This suggests that the weighted model is more appropriate for understanding the relationship between education, trade openness, and protests in Russia. From the weighted model, we see that a weighted model which emphasises tertiary education has the strongest interaction, supporting the idea that higher levels of education can leader to inequality in trade exposure, thus economic grievances arise and increase the likelihood of protests.

Table 6: Negative Binomial Regression: Interaction Effects of Education on Trade Openness and Protests

	Secondary Educ	Tertiary Educ	Summed Educ	Weighted Educ	Non-Linear Educ
(Intercept)	1.864***	2.206***	3.478***	3.042***	−0.751
	(0.446)	(0.594)	(0.813)	(0.802)	(1.710)
Trade Openness	0.060***	−0.025*	−0.063**	−0.048**	−0.007
	(0.011)	(0.012)	(0.022)	(0.017)	(0.023)
Secondary Education	0.001				0.119
	(0.018)				(0.095)
Trade Open × Secondary Education	−0.002***				−0.001+
	(0.001)				(0.001)
Tertiary Education		−0.017			0.121
		(0.023)			(0.094)
Trade Open × Tertiary Education		0.002***			0.002***
		(0.000)			(0.001)
Combined Education			−0.039*		
			(0.016)		
Trade Open × Combined Education			0.002***		
			(0.000)		
Weighted Education				−0.054+	
				(0.032)	
Trade Open × Weighted Education				0.003***	
				(0.001)	
Secondary Education Squared					−0.003
					(0.002)
Tertiary Education Squared					−0.003
					(0.002)
Num.Obs.	237	237	237	237	237
RMSE	28.10	21.09	31.69	22.11	20.50

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

Standard errors are shown inparentheses. The table shows the results of five negative binomial regression models estimating the effect of trade openness on protests, moderated by different levels of education. The 'Weighted Education' model assigns 30% weight to secondary education and 70% to tertiary education.

3.4 Expanding the Model

To further complete the model to better understand the relationship between trade openness and protests, we can include additional control and moderator variables. Control variables can be included to ensure there are no omitted confounders, whereas moderator variables, such as education can be added to explain how the effect of trade openness on protests changes. Therefore, for deciding whether to include additional moderaters, we need to consider whether there are variables which may alter the strength of the relationship between trade openness and protests. One key variable which may alter effects of trade openness on protests is the interaction of population `reg_pop` size which is a proxy for the mobiliasation capacity of the region. We can therefore propose a new model specification to consider this moderator:

$$\begin{aligned} \text{Protest Count} = \beta_0 + \beta_1 \,\text{tradeopen} + \beta_2 \,\text{weightededucation} + \beta_3 \,\text{regionpop} \\ + \beta_4 \,(\text{tradeopen} \cdot \text{weightededucation}) + \beta_5 \,(\text{tradeopen} \cdot \text{regionpop}) + \varepsilon \end{aligned} \tag{9}$$

The results from adding this additional moderator show that the interaction of trade openness and population is statistically significant such that the when population increases, trade openness is more effective at reducing the number of protests. This suggests that the effect of trade openness on protests is moderated by the population size of the region by potentially allowing the benefits of trade to be widespread across the region, or there may be more focus on state control in these dense regions. This moderator should be kept in the analysis. Next, we should include additional control variables. Palmtag, Rommel and Walter (2020) suggest controlling for political variables such as regional welfare, economic grievances, understanding of regional developments, regional structural conditions, and economic globalisation.

A final model is now fitted to include these controls, such that the model shows the effect of trade openness on protest activity, moderated by education and population, while controlling for key socio-economic and demographic factors. The control variables in the model are `grp`, `reg_grpgr`, `reg_levelofunempl`, `reg_avwage`, `reg_urbanshare`, `lnroadden`, and `pctRussian`. The results from this fully specified model are shown in *Table (7)* and are particularly interesting since adding control variables. The results show that there is no longer a significant effect of education as a moderator for trade openness has no effect on protests. Instead, population becomes the primary driver of protest activity. When evaluating the model's coefficients, a unit increase in population (1 million additional people) will double the expected count of protests for a given region, holding all else constant. Trade openness is no longer a direct influence on protests but has a strong effect when interacted with population. Protests are also suppressed where there is higher economic growth; although, perhaps because of the earlier stages of growth (during urbanisation), more protests are seen.

4 Exercise 4: LGBTI acceptance across Africa

4.1 Data Summary

This exercise looks at attitudes towards LGBTI individuals across Africa. In particular, we are interested in understanding the factors determining the attitudes towards LGBTI individuals in Africa. The dataset used

Table 7: Negative Binomial Regression: Interaction Effects of Education and Population on Trade Openness and Protests

	(1)
(Intercept)	1.113
	(0.741)
Trade Openness	0.010
	(0.020)
Weighted Education	−0.022
	(0.030)
Population	0.724***
	(0.091)
Trade Open × Weighted Education	0.000
	(0.001)
Trade Open × Population	−0.005*
	(0.002)
Num.Obs.	237
RMSE	19.42

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Standard errors are shown inparentheses. The table shows the results of a negative binomial regression model estimating the effect of trade openness on protests, moderated by education and population.

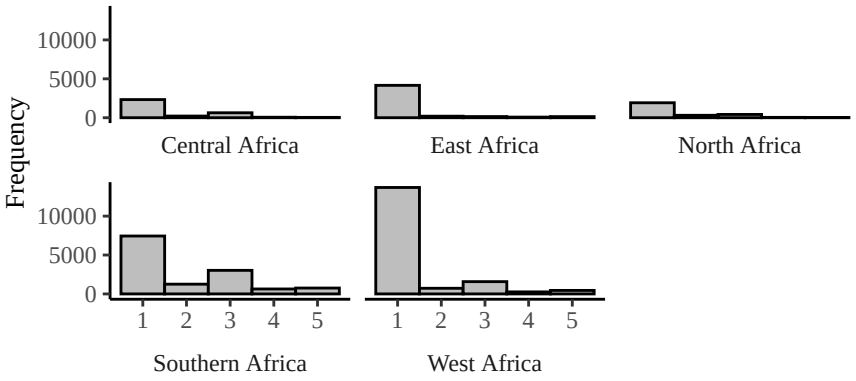
Table 8: Negative Binomial Regression: Impact of Trade Openness, Education, Population, and Controls on Protest Activity

	(1)
(Intercept)	−1.925+
	(1.157)
Trade Openness	0.019
	(0.021)
Weighted Education	0.023
	(0.036)
Population Size	0.718***
	(0.091)
Gross Regional Product (USD Million)	0.000
	(0.000)
GRP Growth (%)	−0.867**
	(0.293)
Unemployment Rate (%)	0.018
	(0.015)
Average Wage	0.000*
	(0.000)
Urbanisation Rate	0.045***
	(0.008)
Distance to Moscow (log)	−0.110
	(0.086)
% Share of Ethnic Russians	0.001
	(0.004)
Trade Open × Weighted Education	0.000
	(0.001)
tradeopen:reg_pop	−0.006*
	(0.003)
Num.Obs.	237
F	21.847
RMSE	18.30

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Standard errors are shown inparentheses. The table shows the results of a negative binomial regression model estimating the effect of trade openness on protests, moderated by education and population, while controlling for ‘grp‘, ‘reg_grpgr‘, ‘reg_levelofunempl‘, ‘reg_avwage‘, ‘reg_urbanshare‘, ‘lnroadden‘, and ‘pctRussian‘

in the analysis, `afrobarometer.csv`, contains data on a wide set of attitudes towards LGBTI individuals in Africa, along with the core dependent variable of interest, `Q87C` which measures attitudes towards having homosexuals as neighbours on a Likert scale of 1 (strongly dislike) to 5 (strongly like). On initial inspection of the `Q87C` variable, *Figure (4)* shows a histogram plot of the views on LGBTI neighours in Africa, grouped by `REGION`, showing that frequency density is skewed towards `1` on the Likert scale, which is particularly pronounced in Southern and Western Africa.

Figure 4: Histogram of Attitudes towards LGBTI Neighbours by African Region



Notes: The histogram shows the distribution of attitudes towards having homosexuals as neighbours (Q87C) across different African regions.

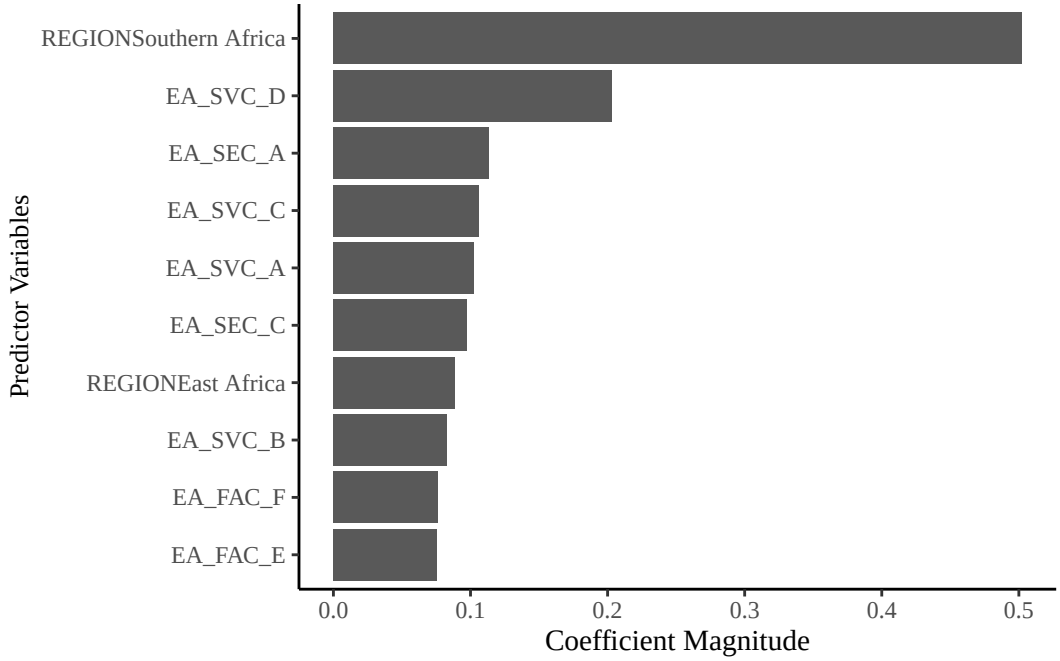
4.2 LASSO Regression Model

This section looks at which variables best predict the attitudes towards LGBTI neighbours by using a LASSO regression model. This model is used to keep only the most important predictors in the model by penalising overfitting. As we have a large number of variables, the model is particularly useful here to identify the best predictors in our model. By splitting the data into 5 folds for training and testing, we get a Mean Squared Error (MSE) plot across different lambda values. The best lambda value (the value which is smallest) is then used to fit the LASSO model. This lambda value is `0.001` and is then used to test the model's performance and identify the best predictors. The results of the LASSO model are shown in *Figure (5)* below. This shows that the the variables which are most important in predicting attitudes towards LGBTI neighbours are `REGIONSouthern Africa`, `EA_SVC_D`, and `EA_SEC_A` which are the region of Southern Africa, access to cell phone service, and police in the area respectively.

4.3 Comparison to Regression Trees

We can also use a regression decision-tree model to predict attitudes towards LGBTI individuals in Africa. We start by splitting the data into smaller and smaller groups based on the values of the predictors. The model then chooses a predictor that best splits the data into two groups (by reducing variability), and continues this process until the criteria (e.g., maximum depth) is reached. As with the LASSO model, we can obtain the top predictors of attitudes towards LGBTI individuals in Africa. The results of the regression tree model are shown in *Figure (6)* below. This shows that the the variables which are most important in

Figure 5: LASSO Regression: Top 10 Predictors of Attitudes towards LGBTI Neighbours in Africa



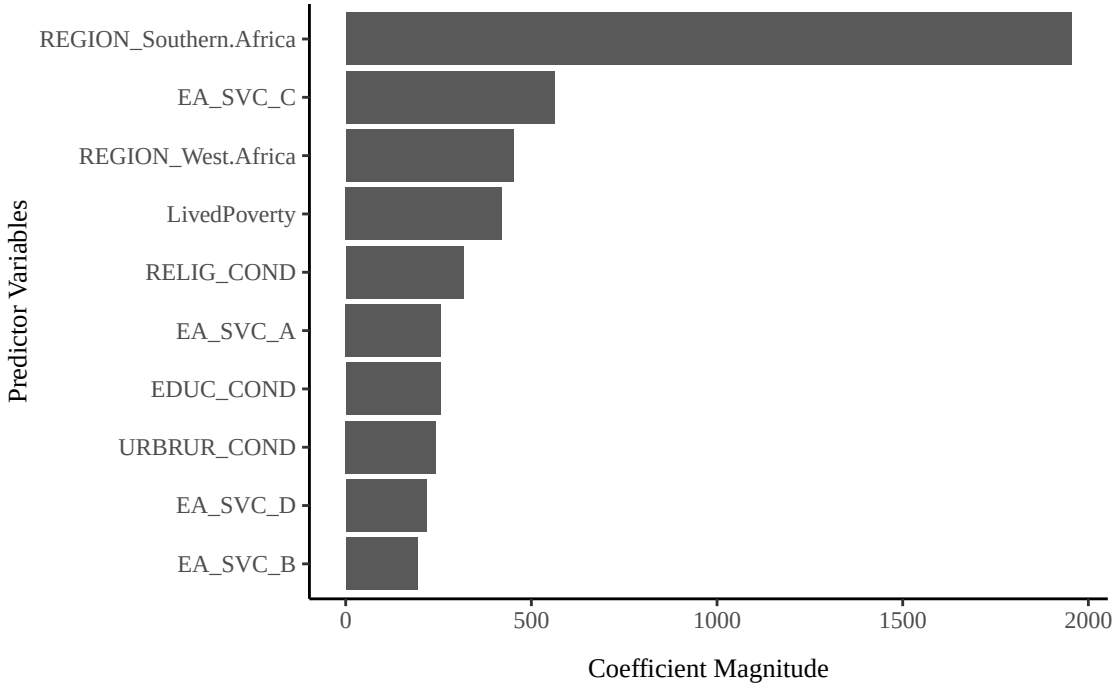
Generated using the LASSO regression model. Variables with higher importance contribute more to the prediction.

predicting attitudes towards LGBTI neighbours are `REGIONSouthern Africa`, `EA_SVC_C`, and `LivedPoverty` which are the region of Southern Africa, has a sewage system for most houses, and the Lived Poverty Index.

Both models emphasise the value of `REGIONSouthern Africa` in predicting attitudes towards LGBTI individuals in Africa. However, due to the nature of the models, they uncover different drivers and complex patterns for prediction. In particular, due to the linear nature of the LASSO model, many interactions may be missed; whereas, the regression tree model can capture these non-linear interactions and capture more complex relationships.

Yet, regression trees can also be problematic. Primarily, they are prone to overfitting meaning noise is often captured rather than the general patters. This can occur when a tree starts going too deep. However, the approach to limit the tree's size of pruning is also problematic. By yielding smaller trees, pruning makes the best splits at each node which is a short-sighted approach that reduces RSS. Moreover, as there are very discrete choices made at each node, any variation in the data can have large effects on the outcomes of the model. Alternatively, a random forest can be used to overcome these limitations. This model is an a set of decision trees which can reduce the variance of the model and improve the prediction accuracy. They work by taking each split in the tree and randomly selecting a subset of predictors to make the split. This means that the model is less likely to overfit as average predictions are created which reduces variance and overfitting of any one tree. This random nature also means they are more robust to variations in the data and the ensemble nature means we can better generalise more stable and accurate data from the random forests.

Figure 6: Random Forest: Top 10 Predictors of Attitudes towards LGBTI Neighbours in Africa



Note: Generated using the regression tree model. Variables with higher importance contribute more to the prediction.

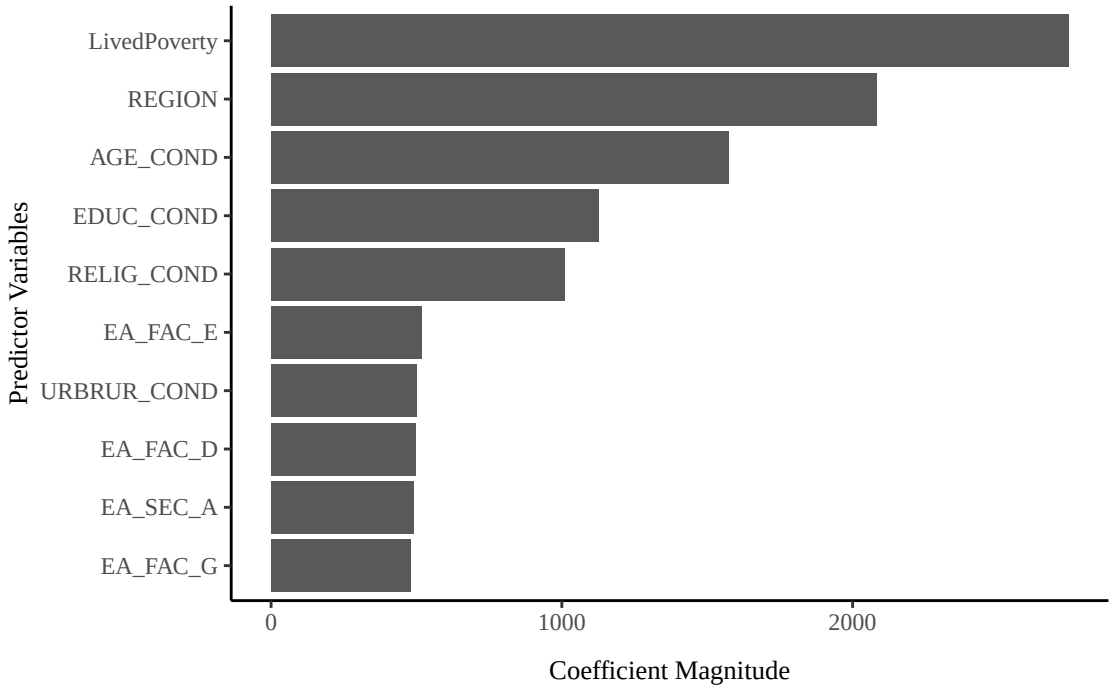
4.4 Random Forest

We finally use a random forest model to predict attitudes towards LGBTI individuals in Africa. The model predicts `LivedPoverty`, `REGION`, and `AGE_COND` which are the Lived Poverty Index, region, and age group respectively. Again, these most important predicters differ from the previous models. By therefore finally estimating the RMSE values for each of the respective models, random forest modelling is the optimal model as it minimises RMSE to `0.968` compared to `0.986` for regression trees and `1.02` for LASSO regression. This suggests that we should use the random forest model to predict attitudes towards LGBTI individuals in Africa as it will have the best accuracy whilst being able to consider the complex relationships. It has the best predictive power whilst akso being robust, despite being harder to interpret.

Bibliography

Palmtag, T., Rommel, T. and Walter, S. (2020) 'International Trade and Public Protest: Evidence from Russian Regions', *International Studies Quarterly*, 64(4), pp. 939–955. Available at: https://doi.org/10.1093/isq/sqaa073.

Figure 7: Random Forest: Top 10 Predictors of Attitudes towards LGBTI Neighbours in Africa



Note: Generated using the Random Forest model. Variables with higher importance contribute more to the prediction.