Project R.E.E.:
Religion, Economics, Education

Andrew Bunnell

Natalie Wellen

Yohan Min

1. Introduction

This project began as a collaboration between three types of data sources. The three authors brought together work from religion, economics, and education. The concept was to tie together three streams of data not normally used in correlative studies. Our broad hypothesis was that religion likely flourishes best in places relatively underdeveloped economically, as well as in places with comparatively low rates of education. This hypothesis was used to produce the following research question for our study: “Can higher tax rates - which generally correlate to countries with high economic development - when combined with high education levels, also correlate with low religiosity?”

The purpose of this study is to show that religion can be measured as a data set, analyzed to make outcome predictions, and assist in international negotiations. Policy formulation benefits with religion as a factor, and the correlations we find, can help policy makers use this variable.

2. Data

This project started with three variables, each describing at least the 146 countries we analyzed. Each team member was in charge of one of the variables. First, Andrew Bunnell brought the importance of religion by country gathered in the Gallup Poll. Next, Natalie Wellen brought tax revenue as a percentage of GDP. Finally, Yohan Min brought the Education Index, a latent variable found in the human development report data using years of education and literacy rates for example. These three variables all came from different sources to create our research question.

We also gathered together control variables. These were tested and used during the linear regression modeling project aspect. For this Population, GDP per Capita, GDP, the continent of the country, and if the countries were developed or not were all gathered together.

Sources

In total we had six different data sources for gathering the variables. Each of the three main variables were scraped from a webpage and then cleaned in Python and one of the control variables, continent. The rest of the control variables had to be downloaded into an excel file before being uploaded into Python and in one case edited in Excel as well.

The importance of religion was scraped off of Wikipedia and based on the results of the 2009 Gallup Poll. The question asked of citizens of each country was “is religion important in your daily life?” There were three possible responses of “yes,” “no,” and “I don’t know.” The variable that we collected and analyzed was the percentage of people in a country that answered “yes.” The countries that had responses listed were 149, and this was the greatest constraint on our data set.

The tax revenue as a percentage of GDP came from the CIA World Factbook. High tax rates are an indicator of economic development, and low tax rates are an indicator of wealth inequality in a country. Thus we chose taxation as the economic indicator. However, there are many forms of taxation, and so revenue collected was used instead. The taxes making up this data include personal and corporate income, value added, and excise taxes, as well as tariffs. The other revenue streams included in these numbers are social contributions including to social security and hospital insurance, and net revenues from public enterprises such as the United States Post Office. This data is for the most recent year available for each country up to and including 2017, but is mostly from 2015. 221 countries are a part of this data set.

The education index was also scraped from the web from Wikipedia. It is a part of the Human Development Index and is published every year by the United Nations. The education index is made up of the average of the mean years of schooling achieved by people over the age of 25 and normalized by 15 years, averaged with the expected years of schooling normalized by 18 years. The normalizations were chosen based on projected maximums for 2025. This data is from the 2015 index. Although this is 6 years after the Gallup Poll where the importance of religion data came from, we do not expect religion to change dramatically in this short amount of time.

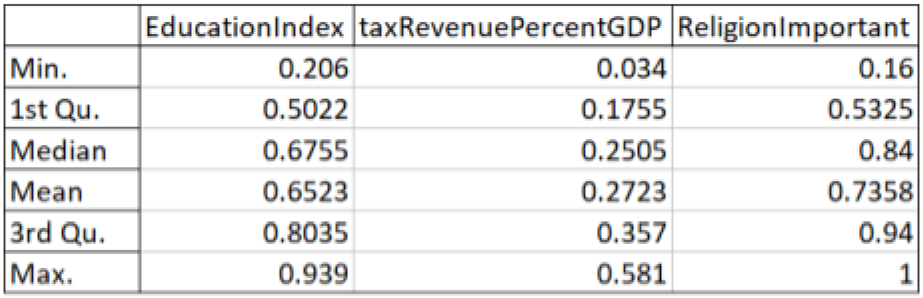
For the control variables there were three different sources. Population, GDP, and GDP per capita in US dollars were all downloaded from the World Data Bank. Population was meant to complement the religion variable, since this is in the form of percentage of population. GDP and GDP per capita were meant to compliment tax revenue as a percentage of GDP. The continents of each country were also scraped off of the web from Statistics Times. The goal here was to ensure we were accounting for underlying geographical patterns that could be taking place. Finally, we also created the binary variable indicating developing countries from the list found on Wikipedia. The list of developing countries came from the International Monetary Fund and was updated in 2018. The list was converted to binary values using if statements in Excel, and then the sheet with the binary variable was uploaded into Python to be merged with the other controls.

Summary of Variables

Our three main variables are all values between 0 and 1. This is good, because being on similar orders of magnitudes means that no variable should have outsized influence while measuring the distance in later computations. Further, the summary in Table 2.1 appears to be of normally distributed variables. This will be tested further on, but this is a good indication for moving forward with the modeling. Importance of religion does vary from this slightly, as the mean of 74% is less than the median 84%.

Further, it worth taking note that there is an inverse relationship between the importance of religion to a population with the average years of education and levels of taxation. This aligns with prior knowledge in the field of international policy and studying the effects of religion. This relationship can be seen clearly in Fig 2.1.

Moving on to the control variables and analyzing our data geographically in Table 2.2, we see that there is a well dispersed representation of countries from around the world. The



	EducationIndex	taxRevenuePercentGDP	ReligionImportant
Min.	0.206	0.034	0.16
1st Qu.	0.5022	0.1755	0.5325
Median	0.6755	0.2505	0.84
Mean	0.6523	0.2723	0.7358
3rd Qu.	0.8035	0.357	0.94
Max.	0.939	0.581	1

Table 2.1 The summary of the three main variables from R.

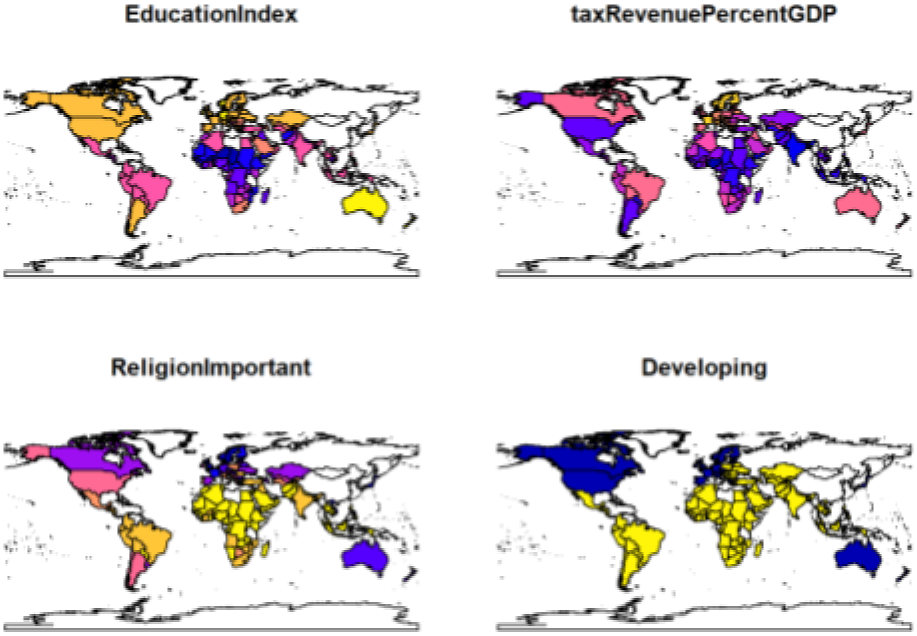
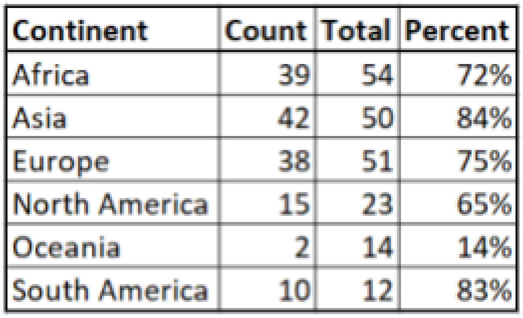


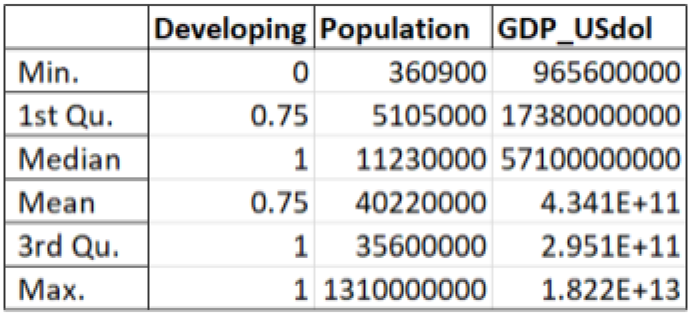
Fig. 2.1 The inverse relationship between importance of religion and the education index and tax revenue is clear from this heat map. As well as the positive correlation between developing countries and importance of religion.

worst case is Oceania, which is made of Australia and New Zealand which are the only two countries that we could get data for. The other countries a part of this continent are very small island nations, so despite the low percentage of countries, we have a high representation from the population. Something similar though less extreme happened in North America as well, where Canada, the United States of America, and Mexico are all included in the data set. Further from examining Table 2.3 we see that 75% of the countries that we examine are considered developing by the International Monetary Fund, and 81% of the countries in the world are developing. So once again we have a good distribution to our data.



Continent	Count	Total	Percent
Africa	39	54	72%
Asia	42	50	84%
Europe	38	51	75%
North America	15	23	65%
Oceania	2	14	14%
South America	10	12	83%

Table 2.2 The number of countries in our data set for each continent.



	Developing	Population	GDP_USdol
Min.	0	360900	965600000
1st Qu.	0.75	5105000	17380000000
Median	1	11230000	57100000000
Mean	0.75	40220000	4.341E+11
3rd Qu.	1	35600000	2.951E+11
Max.	1	1310000000	1.822E+13

Table 2.3 Summary of the numeric control variables from R.

Finally, as we examine population and GDP, our other control variables we see that they are many orders of magnitude greater than our main variables. GDP tends to be at least three orders of magnitude larger than population, so even normalizing this as GDP per capita would still be relatively massive. This is something to keep in mind and pay attention to in analysis of the linear regression models we will look at later.

Integration of Variables

All of the integration of these variables was completed in Python. First, it was ensured that each data set was organized by country so that we had a way to merge. Next, each data set was cleaned of NaN values by searching for symbols the computer would not recognize when converting to numeric values. In the case of continent, our only ordinal variable, a set was created to ensure there were only values we expected.

To merge the data sets, a dirty merge was first performed. Next, the countries that were not matched from the two sets were listed out, and the shorter list used to create a dictionary. After renaming the proper data frame country values with the dictionary created, the two data sets could successfully be merged. Although our smallest data set started with 149 countries, three of these countries were not found in at least one of the other data sets, and so we ended up with 146 countries to analyze.

Once the data was integrated, it was saved as pickle and RDS files to be accessed again in the future. Once saved in a format that R could read, we were prepared to analyze using the unsupervised method of clustering, and the supervised method of linear regression.

3. Clustering

We believe that countries with similarly high levels of education and taxes and an inverse relation to religion will be clustered together. Thus having an idea of the religiosity for one country at a certain economic and education level will give an indication about the other countries at that level. This approach also gives us a chance to analyze outliers for missing aspects in this characterization.

Technique

First we used a DB Scan. The DB scan is often ideal for outliers and groups clusters together by a nearest neighbor algorithm. To start, epsilon was chosen by examining a Nearest Neighbors Distribution Plot seen in Fig 3.1. We can see here that we have a large variance in distances. Choosing a large epsilon of 0.13 to account for this variance still yielded poor results as seen in Fig 3.2. Instead of creating clusters to analyze, the algorithm found most of the countries to be outliers.

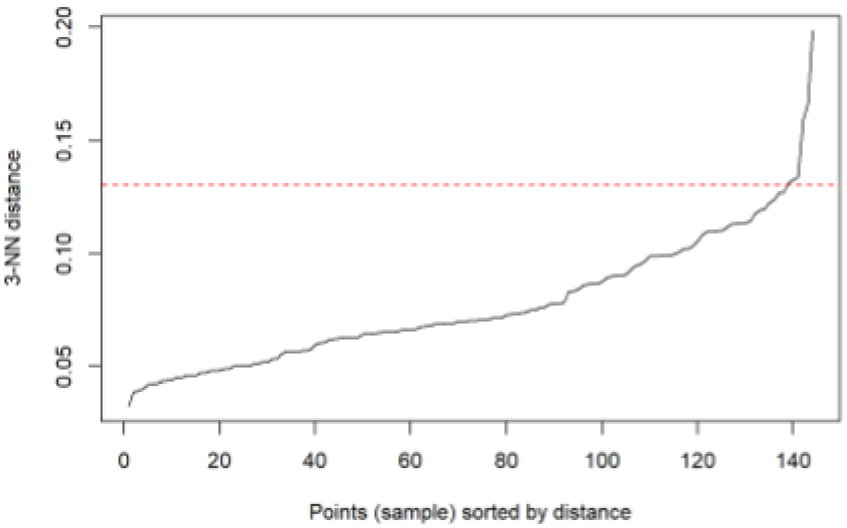


Fig 3.1 The distances of each country from each other sorted by distance. The red dashed horizontal line indicates an epsilon of 0.13 that was chosen to run the dbscan algorithm with.

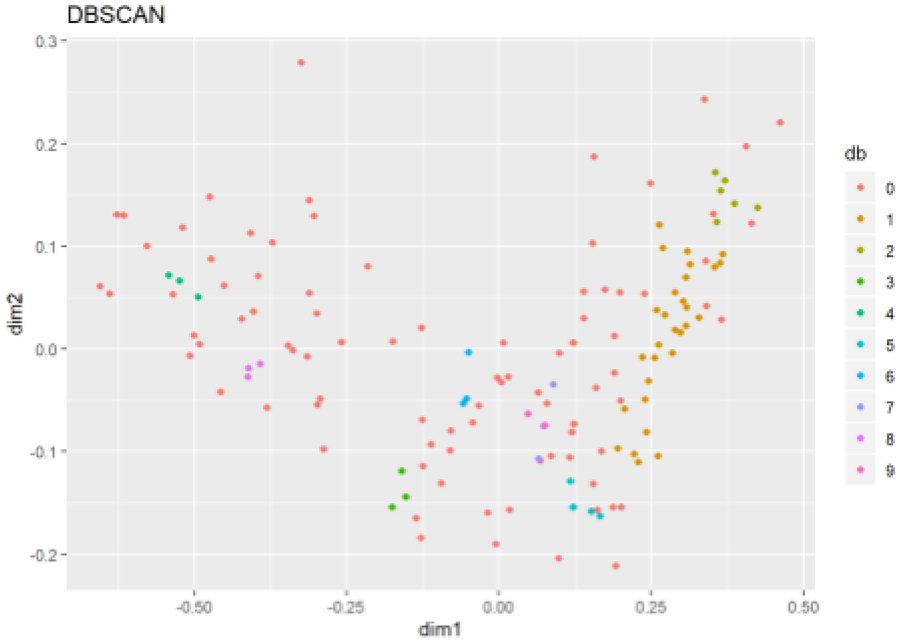


Fig 3.2 Most countries were classified as outliers by DB Scan.

We then ran the PAM, AGNES, and DIANA algorithms.

The PAM algorithm runs in multiple phases. First it organizes the data using medoids, which are similar to taking the average, but must remain a member of the data set. These medoids are then considered selected values, and the number of them is inputted to the algorithm. Then the average dissimilarity of the clusters is minimized to decide which cluster the non-selected members go. Once this initial configuration is created PAM then tries to improve the clusters by randomly swapping the selected node that is being minimized against. This explains why we get the result later discussed where the negative silhouettes are the smallest for the PAM algorithm.

The PAM algorithm was used to determine the number of clusters to use. We tested splitting the data into three, five, or seven clusters. Five clusters were chosen. It did not have as many negative silhouettes as three clusters, which was too coarse for our data. Further it yielded similar negative silhouette results as seven clusters, but seven is much harder to analyze and test without much gain. Thus five was considered the most parsimonious and used as our clustering number to compare the algorithms.

AGNES and Diana are closely related, where AGNES is constructive and DIANA is deconstructive. For AGNES this means that an optimal pairing is made of current groups in each loop of the algorithm. Here each datum starts as a separate group. This is completed until there are only the designated number of clusters so that joining groups again would lead to too few clusters. DIANA goes in the opposite direction, starting with all the data as one group, and optimally splitting one group per loop in the algorithm until there is the designated number.

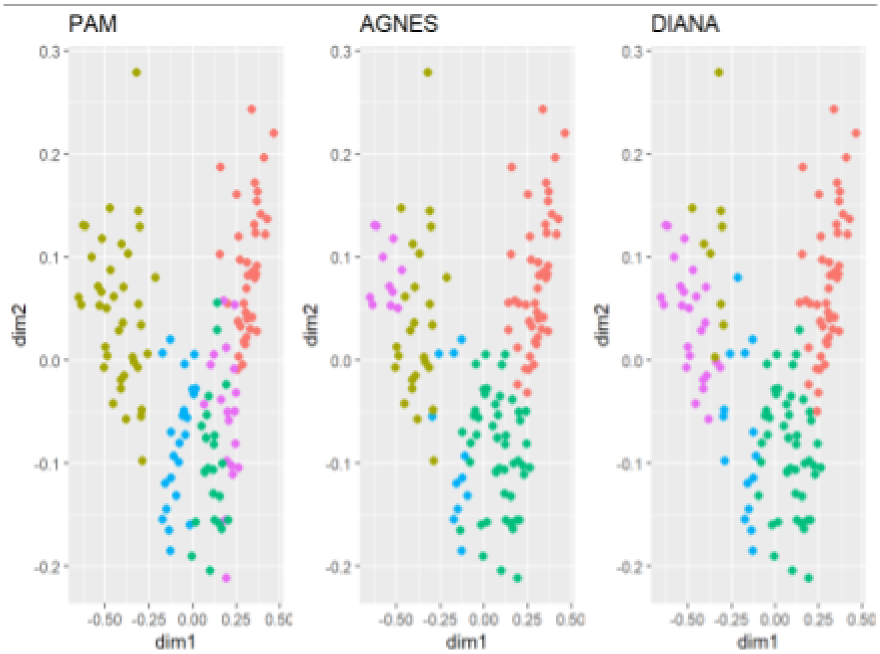


Fig 3.3 The clusters returned by each of the clustering algorithms.

Algorithm Output

The AGNES algorithm we considered the worst output. Clusters two and three yielded nine countries with negative silhouettes. All but one of these was found in cluster three, where the largest negative silhouette width was 0.33 which is 2x larger than the largest returned by PAM. Further, the algorithm results did not fit with any other part of our analysis.

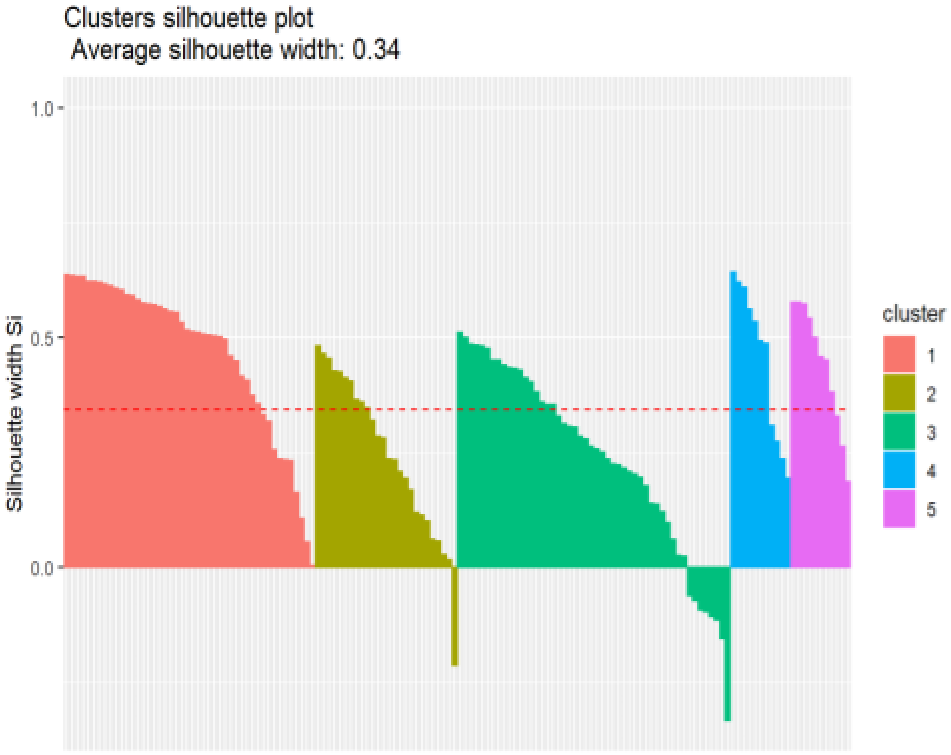


Fig 3.4 The silhouettes returned by AGNES.

Country	Cluster	Sil Width
Netherlands	2	-0.21408656
United States	3	-0.06163778
Kyrgyzstan	3	-0.07235193
Jamaica	3	-0.09294182
Argentina	3	-0.09621792
Malta	3	-0.10643175
Zambia	3	-0.11402789
Moldova	3	-0.15394196
Bosnia&Herzegovina	3	-0.33355791

The DIANA algorithm returned the fewest negative countries. Only seven countries had a negative silhouette. But, the largest negative silhouette was also similarly large like AGNES at -0.3.

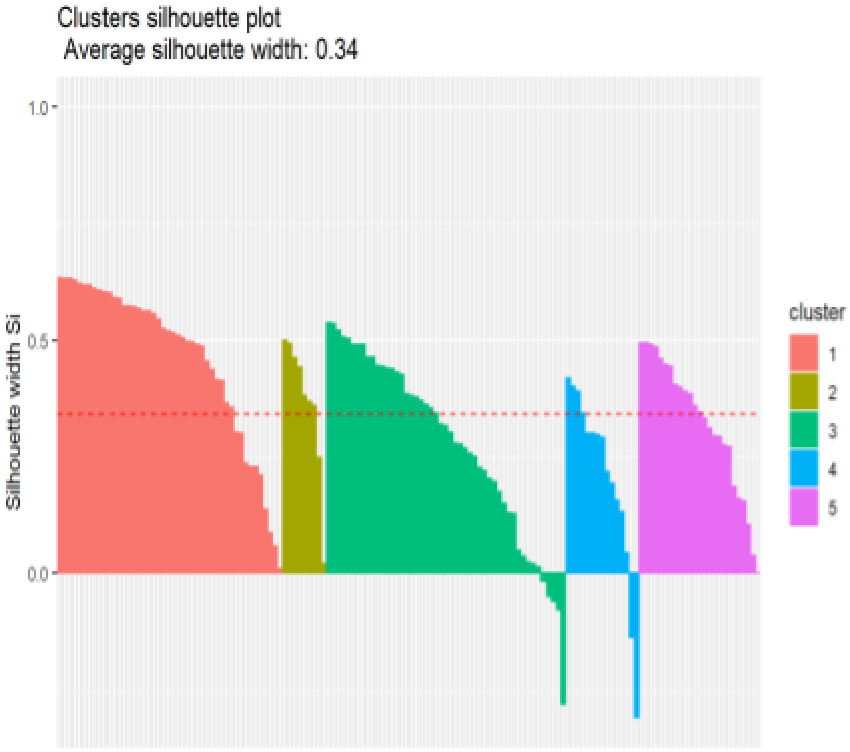


Fig 3.5 The silhouettes returned by DIANA.

Country	Cluster	Sil Width
Moldova	3	-0.01786173
Bosnia and Herzegovina	3	-0.04906460
Singapore	3	-0.06124957
Chile	3	-0.07832357
Cyprus	3	-0.28150732
Israel	4	-0.13806558
Azerbaijan	4	-0.30954291

The PAM algorithm returned ten countries with a negative silhouette width. However, we find that among these anomalous countries, the results are relative. Turkey, for example, shows as a lower taxed and under-educated area in comparison to the European Union. The other anomalies are mostly situated near to higher taxed and higher educated countries. Another possibility is that they have unique factors, such as being a high energy producer with a small population (Azerbaijan), or a small country with a very unique location (Panama). Overall, with PAM we were able to drastically reduce the negative silhouette width and analyzed these clusters.

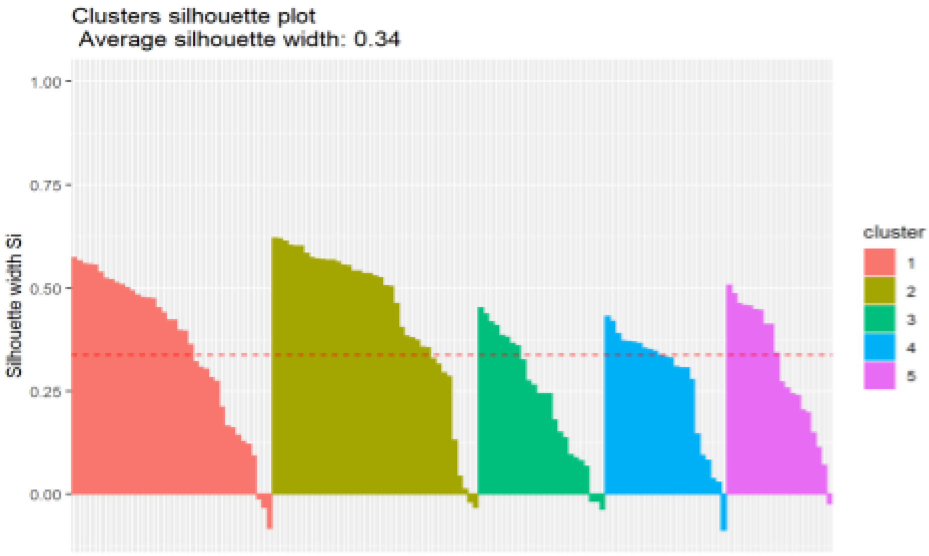
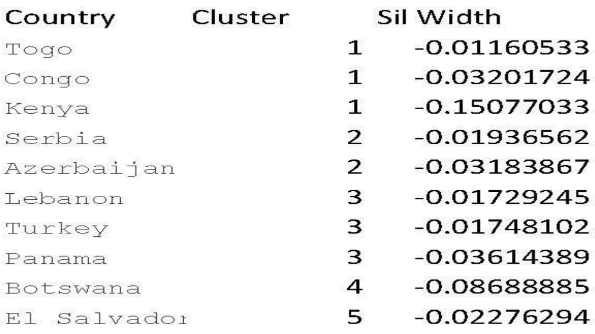


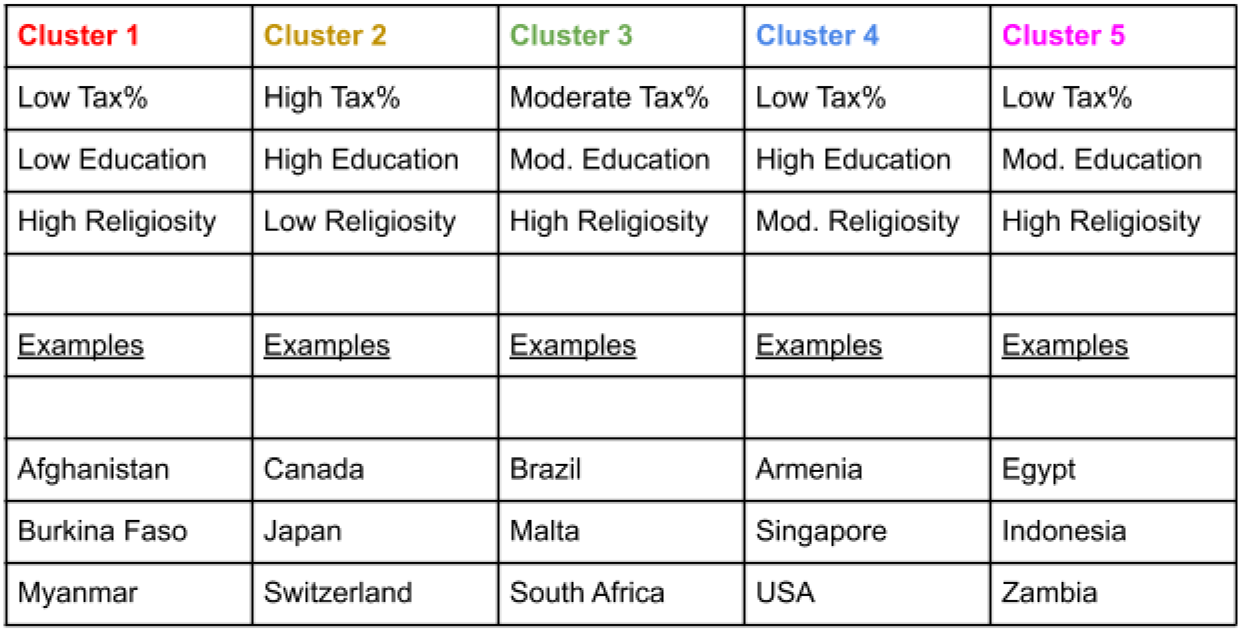
Fig 3.6 The silhouettes returned by PAM.



Country	Cluster	Sil Width
Togo	1	-0.01160533
Congo	1	-0.03201724
Kenya	1	-0.15077033
Serbia	2	-0.01936562
Azerbaijan	2	-0.03183867
Lebanon	3	-0.01729245
Turkey	3	-0.01748102
Panama	3	-0.03614389
Botswana	4	-0.08688885
El Salvador	5	-0.02276294

Results

Countries with high levels of taxation and education and inverse relationship to religion are indeed mostly clustered together. The few countries with negative silhouettes show unique relationships to the three factors. Prevalence of regional religion seems to be a key factor in understanding outliers.



Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Low Tax%	High Tax%	Moderate Tax%	Low Tax%	Low Tax%
Low Education	High Education	Mod. Education	High Education	Mod. Education
High Religiosity	Low Religiosity	High Religiosity	Mod. Religiosity	High Religiosity
Examples	Examples	Examples	Examples	Examples
Afghanistan	Canada	Brazil	Armenia	Egypt
Burkina Faso	Japan	Malta	Singapore	Indonesia
Myanmar	Switzerland	South Africa	USA	Zambia

Table 3.1 Interpretation of the final clusters resulting from PAM.

4. Regression

Hypotheses

We believe that the importance of religion can be predicted by level of education and rate of taxes. A regression will give us a model that can be tested for accuracy on this hypothesis. Here the control variable indicating if a country is developing will also be tested for importance.

Population and GDP, which we considered as control variables, we found insignificant. This is illustrated in Fig 4.1 where the controls are not correlated with “ReligionImport.” We can also see in that figure that religion is inversely related to education and tax revenue as discussed previously in Section 2.

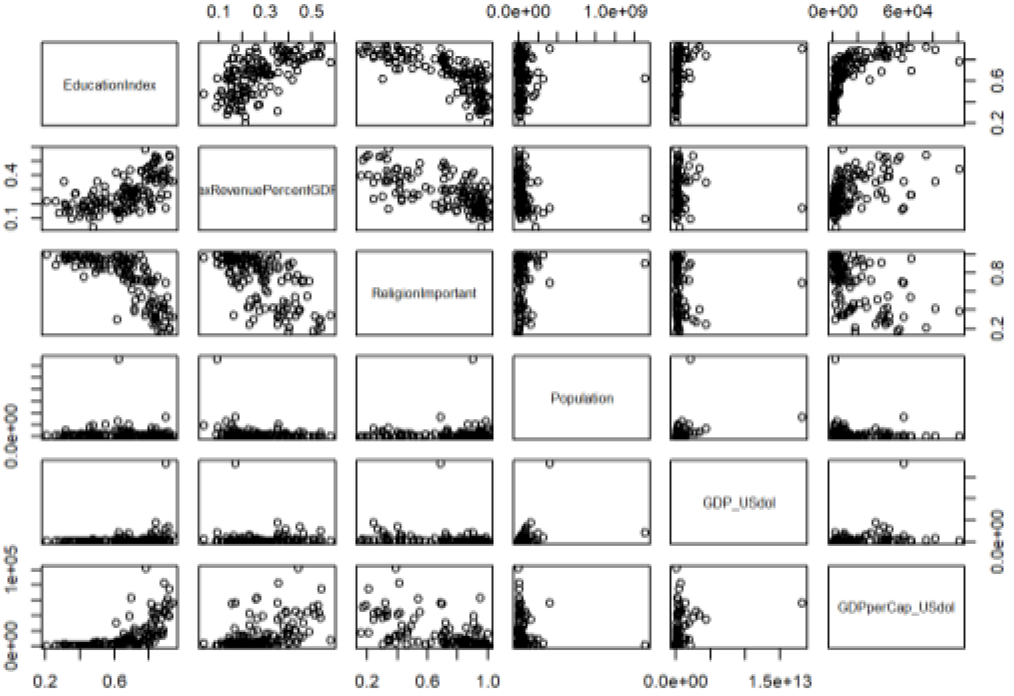


Fig. 4.1 Pairwise correlation plot for the variables

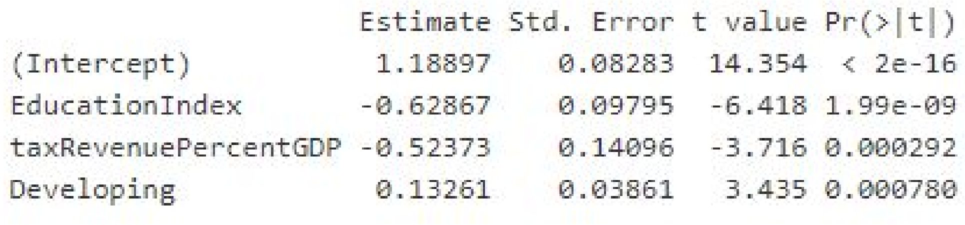
Techniques

We used Gaussian regression and Binomial regression models with control variables. After we found the best model, we tested the model validations to check normality, linearity, homoscedasticity, and influential points (outliers and high-leverage).

Gaussian Linear Regression: Results

We started with the Gaussian Linear Model. After testing different models, we found the best one used the developing country index: 1 if developing, and 0 if developed.

Model: ReligionImportant~ EducationIndex + taxRevenuePercentGDP + Developing



```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.18897    0.08283  14.354  < 2e-16
EducationIndex       -0.62867    0.09795  -6.418 1.99e-09
taxRevenuePercentGDP -0.52373    0.14096  -3.716 0.000292
Developing            0.13261    0.03861   3.435 0.000780
```

Further, education is more a significant predictor of the importance of religion to a country. This was also indicated by our clustering results, matching the two modeling methods.

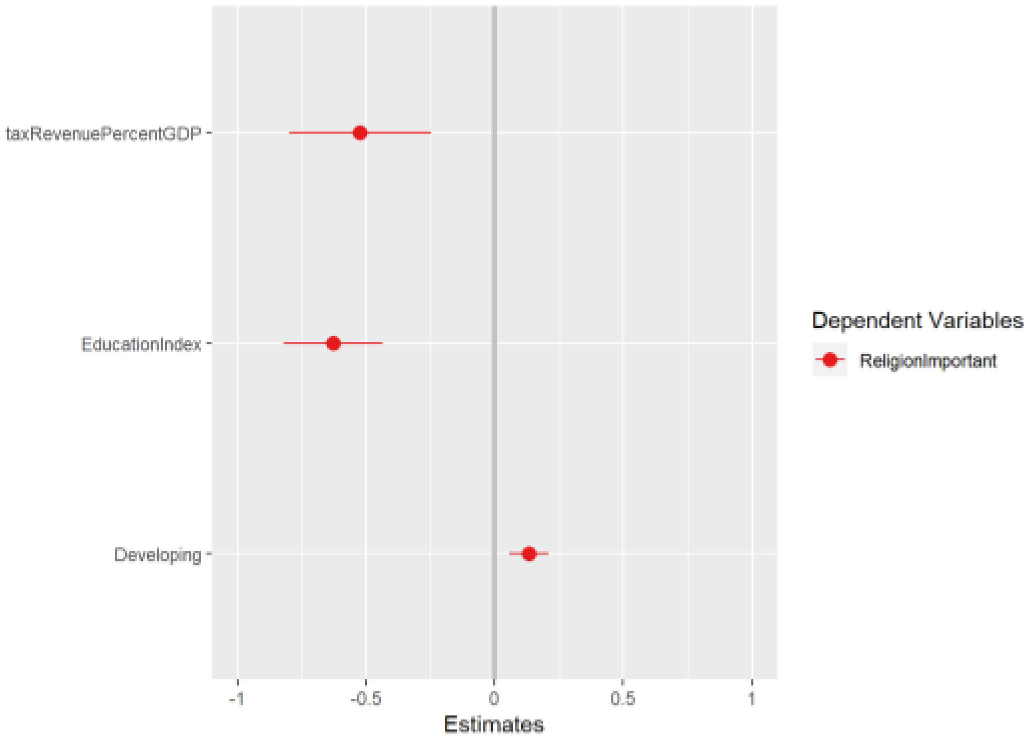


Fig. 4.2 Estimated coefficients of covariates of the regression model

Next we tested this model for validity using a variety of techniques with respect to predictor linearity, normality, heteroscedasticity, outliers, and high leverage points. Along the line of the validity tests, we found there are about three countries that we need to pay attention to which are Viet Nam, Russian Federation, and Hong Kong. Also there are some trends and outliers, in general, the model seems fine to go further with analyses.

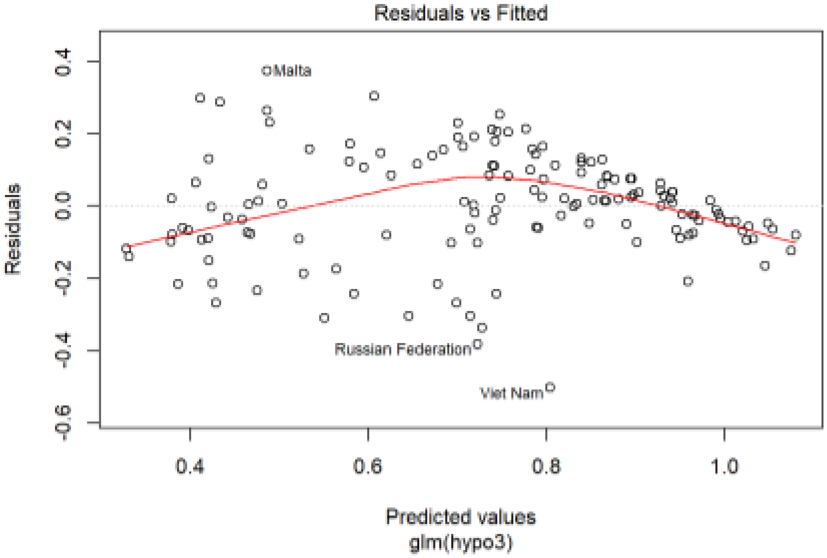


Fig. 4.3 Linearity between dependent variable and predictors

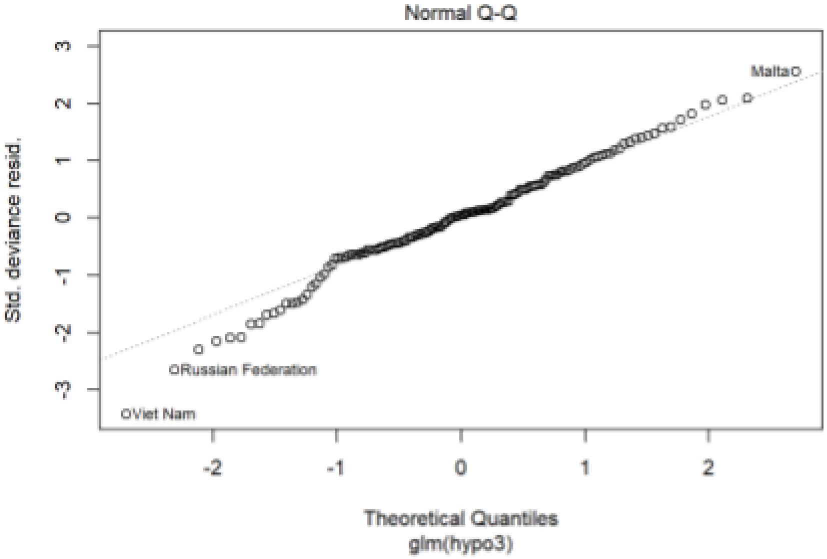


Fig. 4.4 Normality of residuals of the regression model

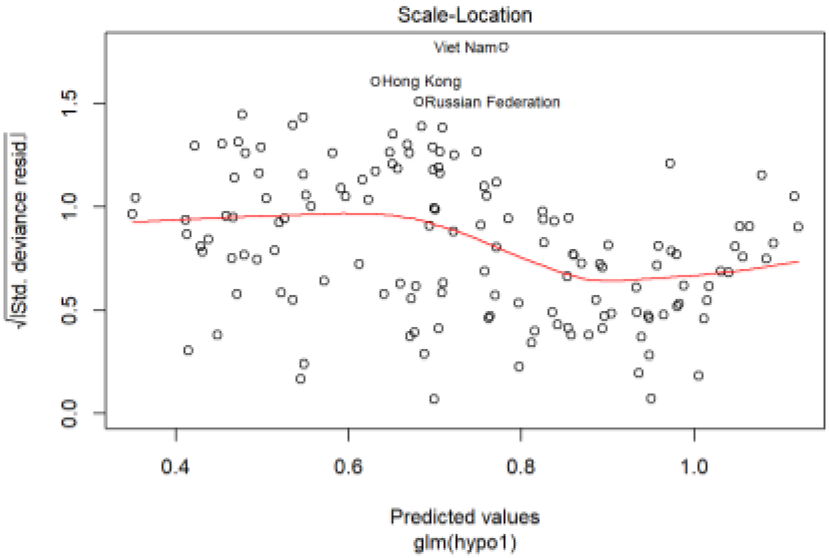


Fig. 4.5 Homoscedasticity of variance of the residual

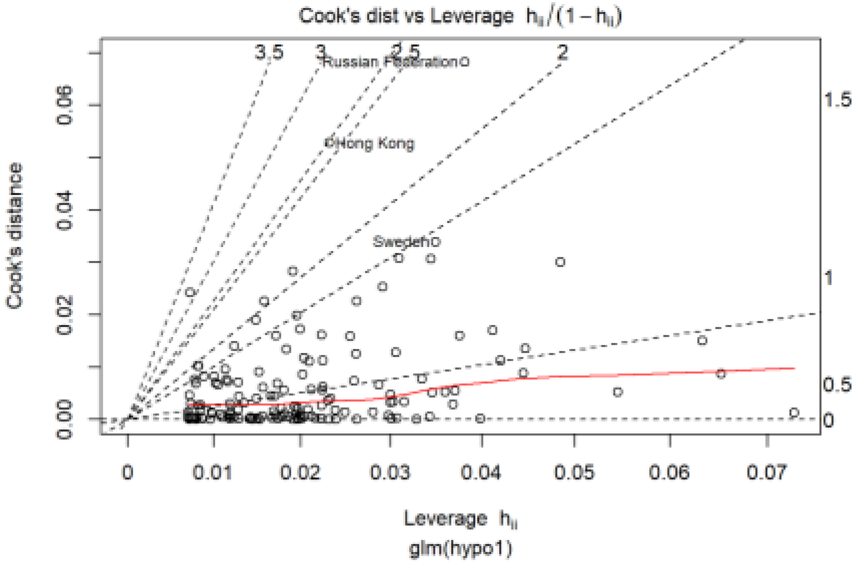


Fig. 4.6 Outliers and high-leverage points of the regression

What we found from this model is that If developed, education and the importance of religion show a linear inverse relationship. If developing however, there is often not much difference in the importance of religion among countries with a lower education index. This can all be seen in Fig 4.7. We interpret this to mean that if the education index is lower than 0.6, those countries consider religion as important regardless of further education distinctions. During clustering, this is where tax revenue became helpful in distinguishing further the importance of religion.

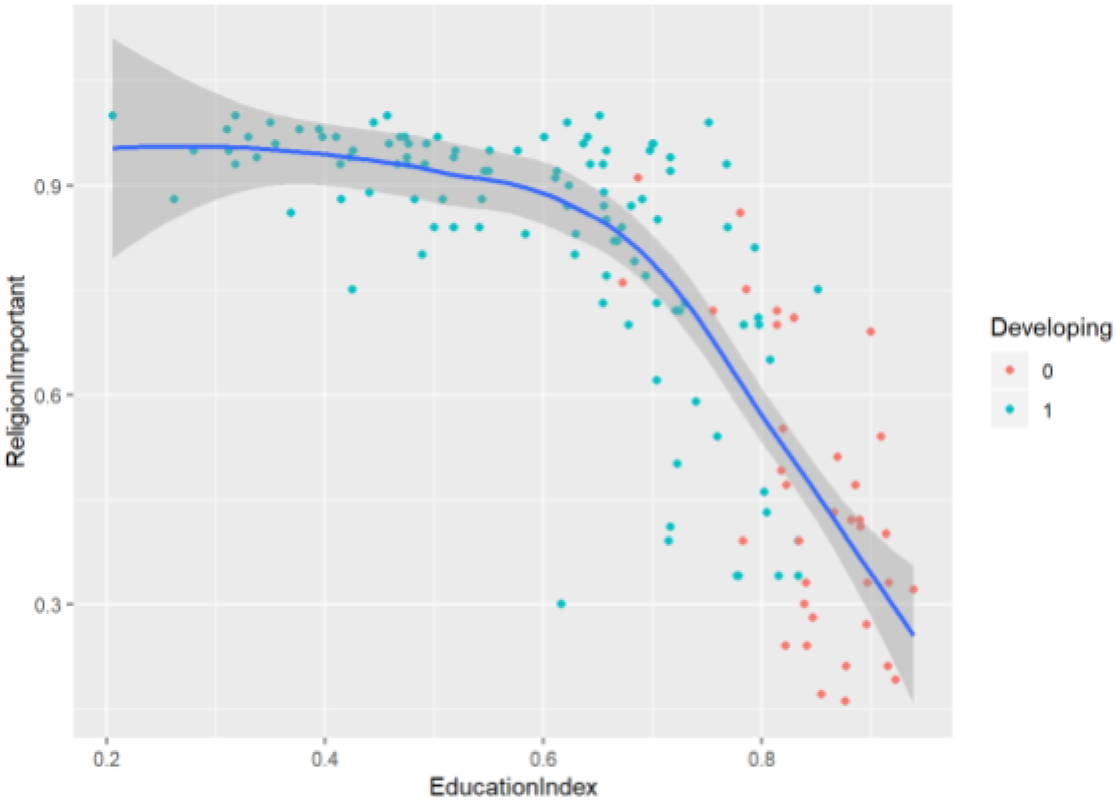
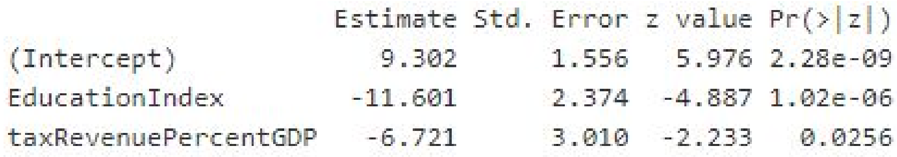


Fig. 4.7 A plot between education and religion variables regarding developing

Binomial Regression Model: Results

Model: Religion (0,1) ~ EducationIndex + taxRevenuePercentGDP

The Binomial Regression Model was once more tested for a different hypothesis with control variables This time however, none of the control variables contributed to lowering the AIC, including the “Developing” variable. Thus, we decided to analyze the model without it.



```
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)             9.302      1.556   5.976 2.28e-09
EducationIndex        -11.601      2.374  -4.887 1.02e-06
taxRevenuePercentGDP   -6.721      3.010  -2.233   0.0256
```

Same as the Gaussian Linear Model, we got the result that the education index has the most significant influence on importance of religion with less variance.

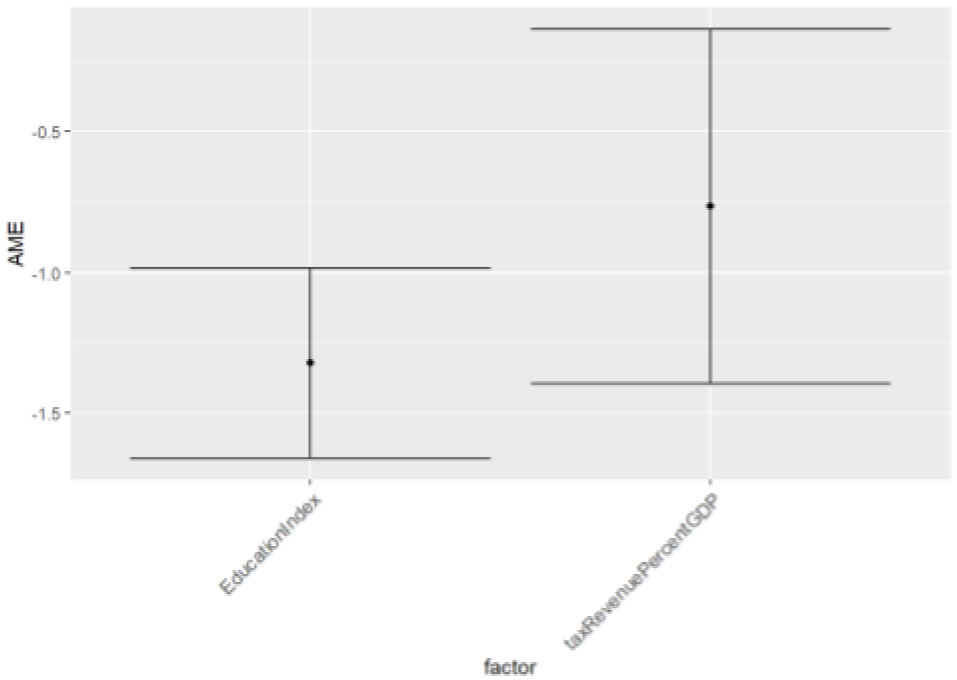
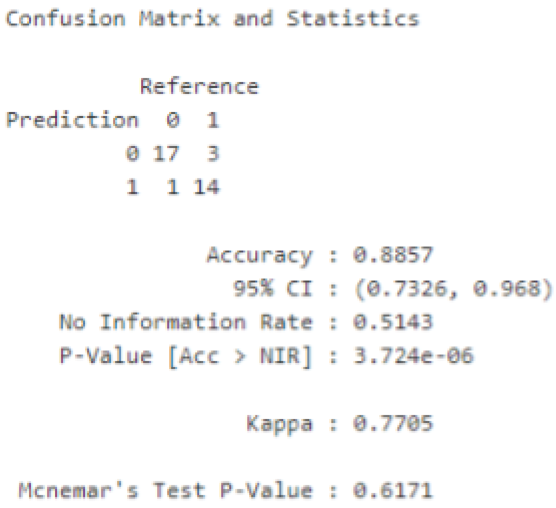


Fig. 4.8 Marginal effect for education and tax revenue variables

We then split the data into train and test sets with the ratio of 0.75 for the binomial regression model above described. We tested the power of the model prediction by using the test data previously split out with cross validation techniques. The results show an accuracy of 0.886. If one does not have any information on a country, then this is a good predictor of the importance of religion in a country. Just like with any data modeling, it is not perfectly accurate. However, this is a rather good result, and supplemental qualitative information about the importance of religion in a country can be used to supplement the Binomial Regression Model results.



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 17  3
         1  1 14

               Accuracy : 0.8857
                 95% CI : (0.7326, 0.968)
    No Information Rate : 0.5143
    P-Value [Acc > NIR] : 3.724e-06

                  Kappa : 0.7705

 Mcnemar's Test P-Value : 0.6171
```

Table 4.1 Confusion matrix

5. Conclusion

We learned that the measure of religiosity in a given country can indeed be predicted from more well known variable values. Since it is unclear where the cause and effect lies, it could be inferred that if a policy goal is to weaken the power of religious institutions in a country, higher tax rates and higher rates of spending on government subsidized education can be an effective way of achieving this outcome. Inversely, lowering the overall tax burden and cutting education spending may increase the relative power of religion in a country. This relationship is something worth exploring further from a public policy perspective.

When policy makers undertake international negotiations for trade deals, climate change talks, and human rights, many background factors are readily studied. Education, gender, health, economics, race, and military spending are examples of important data available to policy makers for analysis when preparing for negotiations. Importance of religion in a given country is a key variable which ought to be used alongside these other ones to help understand the way specific countries look at other social factors. In this sense, policy makers can use religiosity as a measure to better understand the specific views, biases, and approaches that a given country may bring to the negotiating table.

Finally, a hypothesis for future study, is that religion may fill in the gaps in social care. When governments are providing a higher percentage of social needs, religious institutions lose influence and social capital in a society. Fundamentally, we believe that we have shown that religiosity in a country is a data measurement that should be considered alongside other factors such as tax rates, and education spending.