MADA Project: COVID-19 Vaccinations and recovery

Dawson Dobash

2021-12-10

# 1 Abstract

COVID-19 has been the talk of the world for the past two years, and it is still prevalent. COVID-19 not only affected individual people, but it also affected businesses, health protocols, and the economy of governments. All countries faced this impact no matter their location or how developed their country is. It is truly indeed a pandemic.

I am interested in looking at the difference between the proportion of those who recovered to those who have died for each country in the world. The variables I am considering are: percent of cases within a population, continent the country is on, number of tests per person, gdp per capita, proportion of those vaccinated, government type, and health care funds. By looking at some simple linear regressions and other exploratory analysis, the fit of the model to the data were not good for each simple linear regression. I decided to do a step-wise selection process, and found that the percent of cases and the continent the country resides in has an association to the difference in proportions. However, the fit is still not good at all.

The higher the percent of cases there is in a country, the higher recovery proportion there is when keeping the continent constant. North American countries as a whole has a higher death proportion compared to African countries as a whole when keeping the percent of cases constant.

I am also looking at how the proportion of people vaccinated in each country is affected by multiple factors. The predictors I will be considering are: percent of cases within a population, continent the country is on, number of tests per person, gdp per capita, proportion of those who died, government type, and health care funds.

While looking at vaccination proportions, I decided to do some analyses using training/test data sets and applying them to find the RMSE of the full model and a null model. When finding the RMSE, I found that the full model does a better job in fitting the data compared to the null model. This is a good sign in showing that some combination of predictors have an association to the proportion of vaccinated people for each country.

The next analysis was using a machine learning modeling technique called the LASSO model. This type of modeling will create multiple models from a five-fold cross-validation and a tuning grid which will try to find the best performing model. From the LASSO, we find that model 27 gives the best model, and the percentage of cases in a country, which continent the country resides in, and the type of government have impacts to the vaccination proportions. Compared to Africa, South America and Asia increases the proportion of vaccinations the most while an increase of cases means more vaccinations as well. A monarchy government compared to a republic government increases the proportion of people vaccinated.

# 2 Introduction

## 2.1 General Background Information

COVID-19 hit the world by storm and is a possible ever-lasting change to the way society acts. It also had an impact on each country’s funding towards health care. Actions within the past two years have been taken to help fight against this virus. Although COVID-19 has been here for over two years, there are still problems with under staffing in hospitals and not enough beds at some hospitals. Recovery rates are still high in the United States but what about the whole world. Other countries are still getting hit hard by this virus. In terms of vaccination output, there has been back and forth movements on people getting the vaccination. A poll, close to when vaccinations were starting to disperse, showed that only 51% of adults in the United States said they would get a vaccine (Choi, 2021). Although this happens in developed countries, what about the less developed countries and regions such as Africa. While people in the United States are fighting about whether the vaccine should be given or not, the people in these other countries might have a limited supply. According to the United Nations in September of 2021, there is still an unequal amount of vaccinations going to developed and less developed countries. This could be due to the cost of each vaccination (“COVID vaccines,” 2021). In a paper by Choi, as of January 2021, cases were fairly even between developed and less developed countries (Choi, 2021). This just shows that although COVID-19 is affected everyone, the supply is not equal between all countries. In a paper by Mathieu et al., they show a world map and the distribution of vaccines by country as of April of 2021. We can see that Africa is the continent that has the lowest distribution of vaccines (Mathieu et al., 2021).

One of the goals of this analysis is to look at the difference in proportions of recovered patients and those who have died. This will be beneficial in determining if more care and funding is needed for countries. According to paper by Liang et al., the mortality rates were negatively impacted by the number of tests performed and also based on the countries economic status and government (Liang, Tseng, Ho, & Wu, 2020).

My other goal for this project is to analyze how the percentage of people vaccinated is affected. It could possibly lead to important results that might show that we need more either distribution or requirements of the vaccines. As mentioned above, there is a uneven distribution of vaccinations but is this significantly different and what other factors could affect vaccination numbers?

## 2.2 Description of data and data source

The data I am using is from a mixture of locations. Two sources are from kaggle.com, one from the world bank database, one created by myself based on information from wikipedia, and the other is from the Organisation for Economic Co-operation and Development (OECD) website. My data consists of ten data sets with information over each country in the world (for those that data was collected for). Six of these data sets provide the same information just for different countries where each data set is a different region. There is a data set that includes all of the countries but I want to include a region indicator variable for each country. These data sets includes total covid-19 cases, active cases, total deaths, total recovered, total tests, and population size. This data was updated on kaggle as of September 16, 2021. Another data set I am using is a vaccination data set for each country. This is also data as of September 16, 2021. The other data sets are providing statistics for each country that was not included in the larger data set but I still wanted to use. These data sets include the following statistics/demographics: healthcare funding for 2019-2020, vaccination numbers/percentages, gdp per capita, and government type.

I combined the data sets into one using the countries name as the merging variable. If there are missing data within the variables used in analysis, I will exclude those countries. There will be a lot of cleaning for this data just to make it as I want it.

The variables I will be using in my analysis are:

* **prop\_diff\_recov\_vs\_death** - proportion of those who have recovered versus those who have died from COVID-19 (for those who has gotten COVID-19)
* **Value** - health care funds (in U.S. dollars)
* **prop\_vacc** - proportion of those who are vaccinated (for the whole population of a country)
* **pct\_cases** - percent of the population in each country that has gotten COVID-19
* **location** - the continent in which the country resides in
* **test\_per\_person** - the number of tests a person has taken (on average)
* **gdp\_per\_capita** - the gdp per capita for each country (in U.S. dollars)
* **prop\_death** - proportion of those who have died from COVID-19
* **government** - Type of government (Republic, Monarchy, or Provisional)

Links to the locations I received the data are below:

<https://data.oecd.org/healthres/health-spending.htm>

<https://www.kaggle.com/iamsouravbanerjee/covid19-dataset-world-and-continent-wise?select=Covid+Data+-+World.csv>

<https://www.kaggle.com/anandhuh/latest-worldwide-vaccine-data>

<https://data.worldbank.org/indicator/NY.GDP.PCAP.CD>

<https://en.wikipedia.org/wiki/List_of_countries_by_system_of_government>

## 2.3 Questions/Hypotheses to be addressed

1. Since this data is cumulative and does not break down the amount of cases per month per country, I want to compare the recovery proportion to the death proportion. I want to see if this difference is based on variables such as the proportion of those vaccinated, the amount of funding the health care system has, the amount of tests taken per person, the location of the country, the percent of cases in a country, the gdp per capita, and the government type. This will hopefully show if the amount of health care support really helps patients recover.
2. I will also look at how the percent of people vaccinated for each country is affected by multiple predictors (location of country, tests per person, government type, proportion of deaths, gdp per capita, and percent of cases per country). I am planning on seeing that more developed countries based on GDP per capita will have a more vaccinated population.

Some confounders that I could run into are:

1. Developed countries might have more cases as more travelers visit these countries and it could have an effect on recovery numbers.
2. Unknown dates of when a country received vaccines and amount. This could skew the amount of vaccinations a country receives.

# 3 Methods and Results

## 3.1 Data import and cleaning

For data importing and cleaning, I first had to import ten different data sets into R from multiple sources specified in the description of data and data source section. After this importation, the first thing to do before combining the data set was to create indicator variables for each country. After this, I wanted to make sure that each variable was the right class as some of the numeric variables were character variables. This will help in making sure that each variable was able to be used in the right way. I then was able to combine each continent data set into one world data set to make the importing go easier. After combining the continent data sets, I need to only include the variables from the world data set needed for analysis and make sure that all missing values were removed as I only want full observations. However this would delete some countries that were large in population such as China and the Philippines as they did not have vaccine information.

The next part was looking at all of the other datasets which include the government type, gdp per capita, health care funds, and vaccine datasets. To combine these with the world data set, there was a huge problem that goes with using data sets from different sources. This problem is how to combine the data sets with a specific variable. Fortunately, each data set had a country variable to combine them with. However, the country names for all datasets were different. One included a three letter code for countries and another one used different abbreviations/names for countries. I had to use the ‘countrynames’ package to help change into names I wanted for the three-letter code. Since I could not change all of the countries names to the way I wanted it, I had to unfortunately go into the raw data csv files to change them manually for some of the continent specific data sets. I also used ifelse statements to change the names for the gdp per capita and health care funds data sets. I was able to do these in r because there were not as many country names to change compared to the other data sets I had to manually change. A comment within the proccessingscript.R tells of the countries that needed manual changes. After all of this was good to go, I used the merge function to combine all of the data sets the way I wanted to.

I created some variables from the pre-existing ones that I will use within my analysis. Some of these changes were, for example, creating proportions from percentages and combining two variables. These variables can described above in the description of data and data source section.

One disclosure I have for this is that one data set, health care funds, only had 48 countries when I cleaned the data. Therefore, there are a lot of missing values for only the specific variables that came from the health care funds data set in the merged one.

To look at a more descriptive explanation of all of the steps I took in data cleaning and importing, go to the processing script where it has comments for every line of code.

## 3.2 Question 1: Difference in proportions of recovery and deaths

### 3.2.1 Methods

For question 1’s analysis, I first decided to do some simple exploratory analysis on the data. This included some simple linear regression models just to see how each variable fits with my outcome. After fitting most of the predictors to the outcome, I decided to look at a step-wise procedure which will hopefully create the best model for this data. Before fitting the variables into step-wise selection, I did a correlation matrix to see if some variables are very correlated. After considering the correlation matrix, the variables I considered in my analysis were the proportion of those vaccinated, tests per person, continent the country resides in, gdp per capita, government type, and percent of cases in a country as predictors.

I used the caret package within R to do this step-wise procedure. After doing the step-wise procedure and finding which variables will stay in the model, I created that model just to see how well the model actual is. I created a predicted versus observed values scatterplot and a supplementary residual plot as well. These plots and the model model will help give me an idea of what variables are useful in predicting the difference in proportions of recovered people and people who have died.

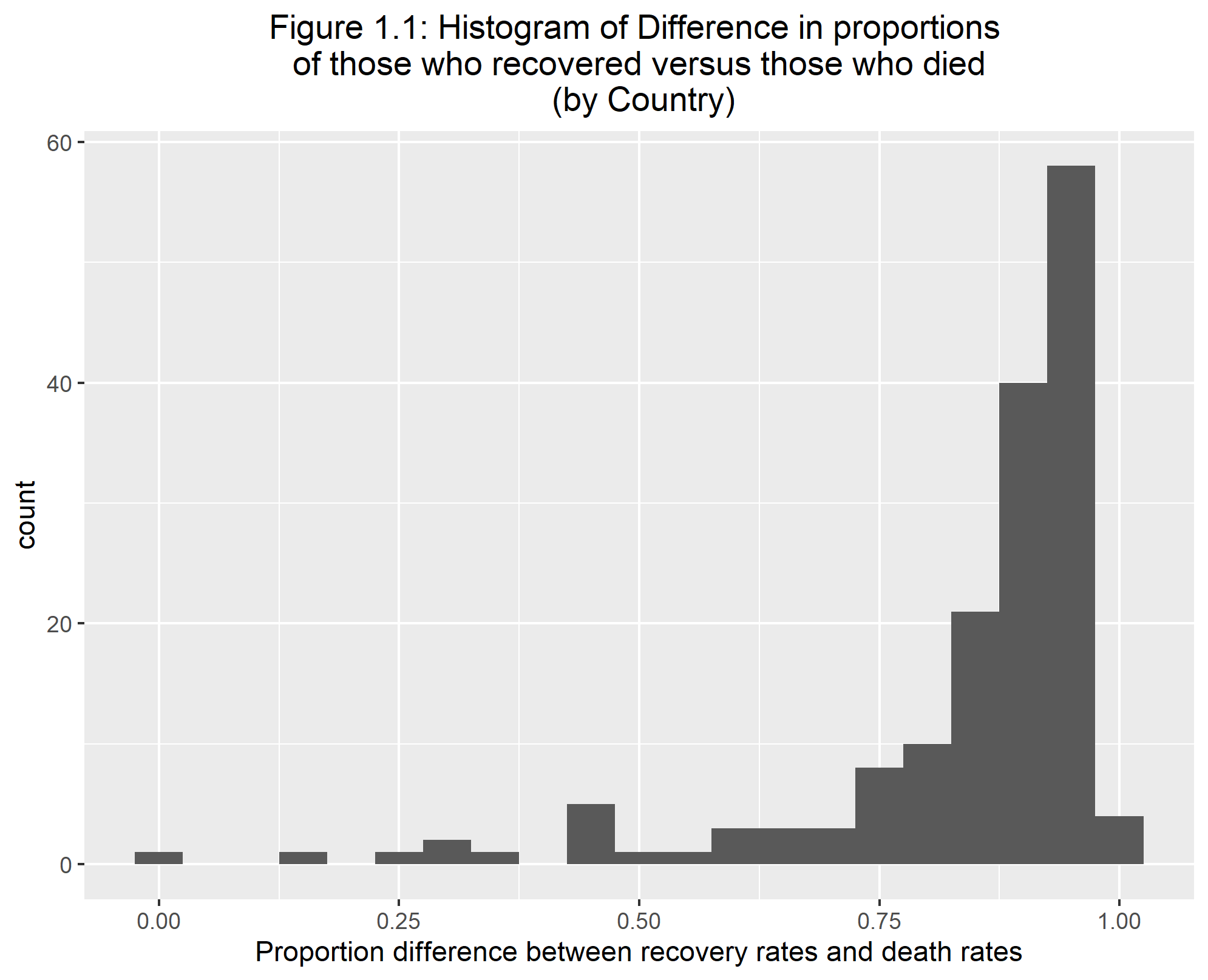
### 3.2.2 Exploratory Analysis

Table 1.1 displays the top five and bottom five death rates. Also within this, it shows their corresponding recovery rates as well. We can see that the bottom five death rates are from Asia but also the highest, Yemen, is within the Asia continent as well. Yemen has a much higher death rate compare to the other top countries. It is more than double the second highest. However, the sample size of total COVID-19 cases is smaller than others.

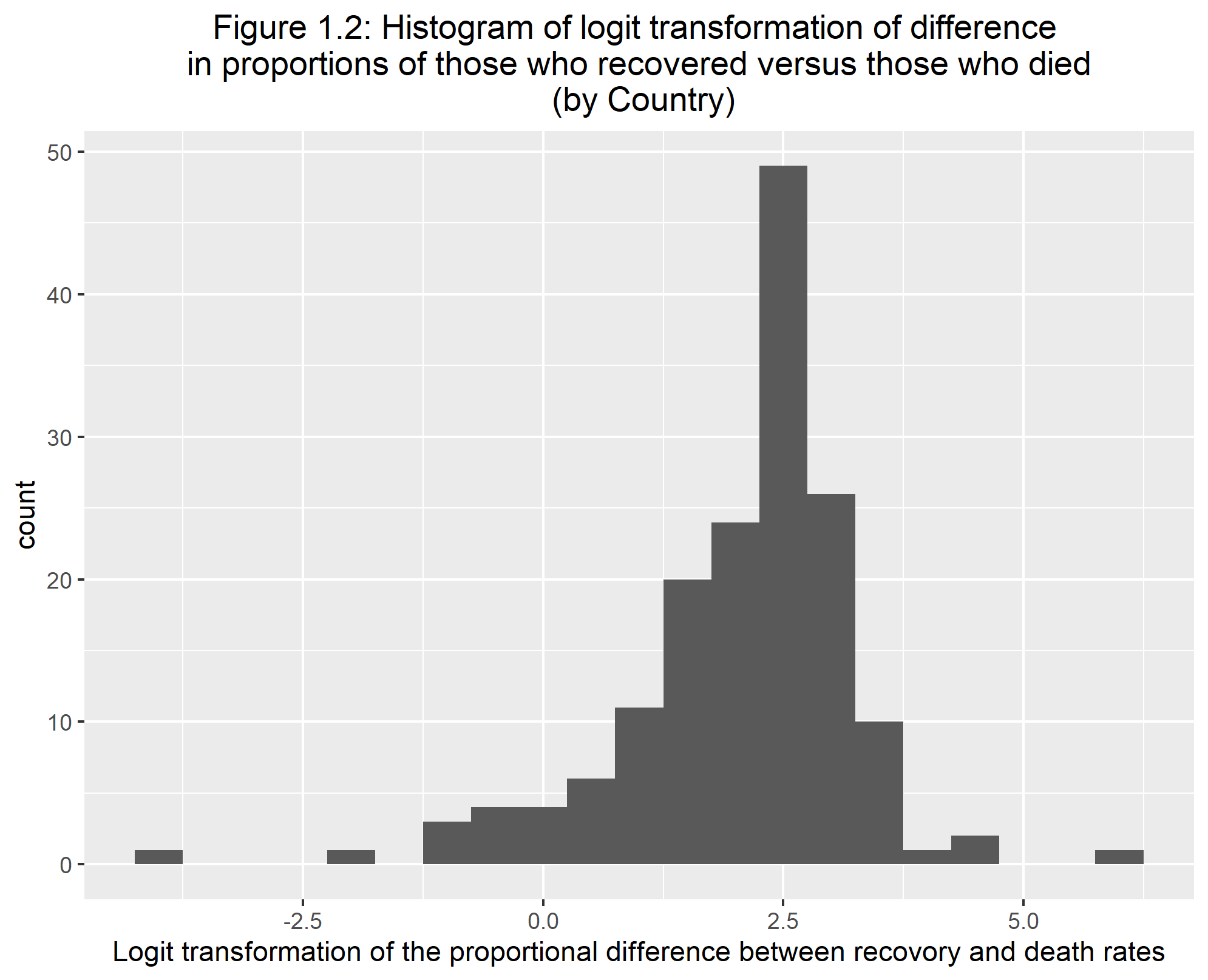
Table 1.1: Top 5 and bottom 5 death rates by country

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Country | location | recov\_rate | death\_rate | Total Cases |
| 1 | Yemen | Asia | 0.6222975 | 0.1892018 | 8557 |
| 2 | Mexico | N. America | 0.8164215 | 0.0761703 | 3549229 |
| 3 | Sudan | Africa | 0.8398947 | 0.0756579 | 38000 |
| 4 | Syria | Asia | 0.7602892 | 0.0697775 | 30153 |
| 5 | Ecuador | S. America | 0.8774760 | 0.0643637 | 505860 |
| 159 | Maldives | Asia | 0.9758072 | 0.0027268 | 83248 |
| 160 | Qatar | Asia | 0.9899229 | 0.0025682 | 235187 |
| 161 | Bhutan | Asia | 0.9984592 | 0.0011556 | 2596 |
| 162 | Laos | Asia | 0.3083227 | 0.0008860 | 18059 |
| 163 | Singapore | Asia | 0.9252218 | 0.0007883 | 74848 |

Figure 1.1 shows a histogram figure that describes the difference in proportions between those who recovered versus the proportion of those who died. From the histogram we can see that this difference is left skewed as most countries have a much higher recovery rate with only a some having more of a lower difference meaning lower recovery rates. This is very worrisome for the analysis as most of the countries are over 0.80 in the difference.



We can see in figure 1.1 that this data is fairly skewed. Therefore, I will do a logit transformation of the outcome to make it more normal as shown in figure 1.2.



Before I decided on the logit transformation, I first decided to try this difference in proportions outcome I made within a binomial model and then a quasibinomial using the values of total cases as my weight to see how the deviance and dispersion values are. After running two simple models, the results were showing very high values for these statistics. Therefore this is how I decided on using the logit transformation.

Using the logit transformation, I ran 4 simple linear regression models. Below I provide the statistics of one of the models within Tables 1.2. Table 1.2 is showing statistics for the predictor vaccine percentage.

Table 1.2: Summary statistics for linear regression (logit transformation of prop diff of recovery and deaths vs. vaccine)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.0403567 | 0.0343962 | 1.184012 | 6.770678 | 0.0101312 | 1 | -257.8129 | 521.6257 | 530.907 | 225.7033 | 161 | 163 |

Looking at Table 1.2, we can see that the R^2 values are very low, and the deviance is also somewhat high. This is a similar situation to all of the other simple linear regression models. The statistics for the other three models are within the supplemental material. Therefore none of these models are good fits to the data. I am going to see if a step-wise selection of predictors could increase the fit of the model to the data.

### 3.2.3 Main Analysis

The main analysis pertains to using step-wise selection for the variables I specified within the methods section for question 1. I am doing step-wise selection because for each simple linear regression, the r-squared values are very small and I wonder how my model will do if I try to select the best variables in predicting the logit transformation of the difference in proportions between recovery and deaths.

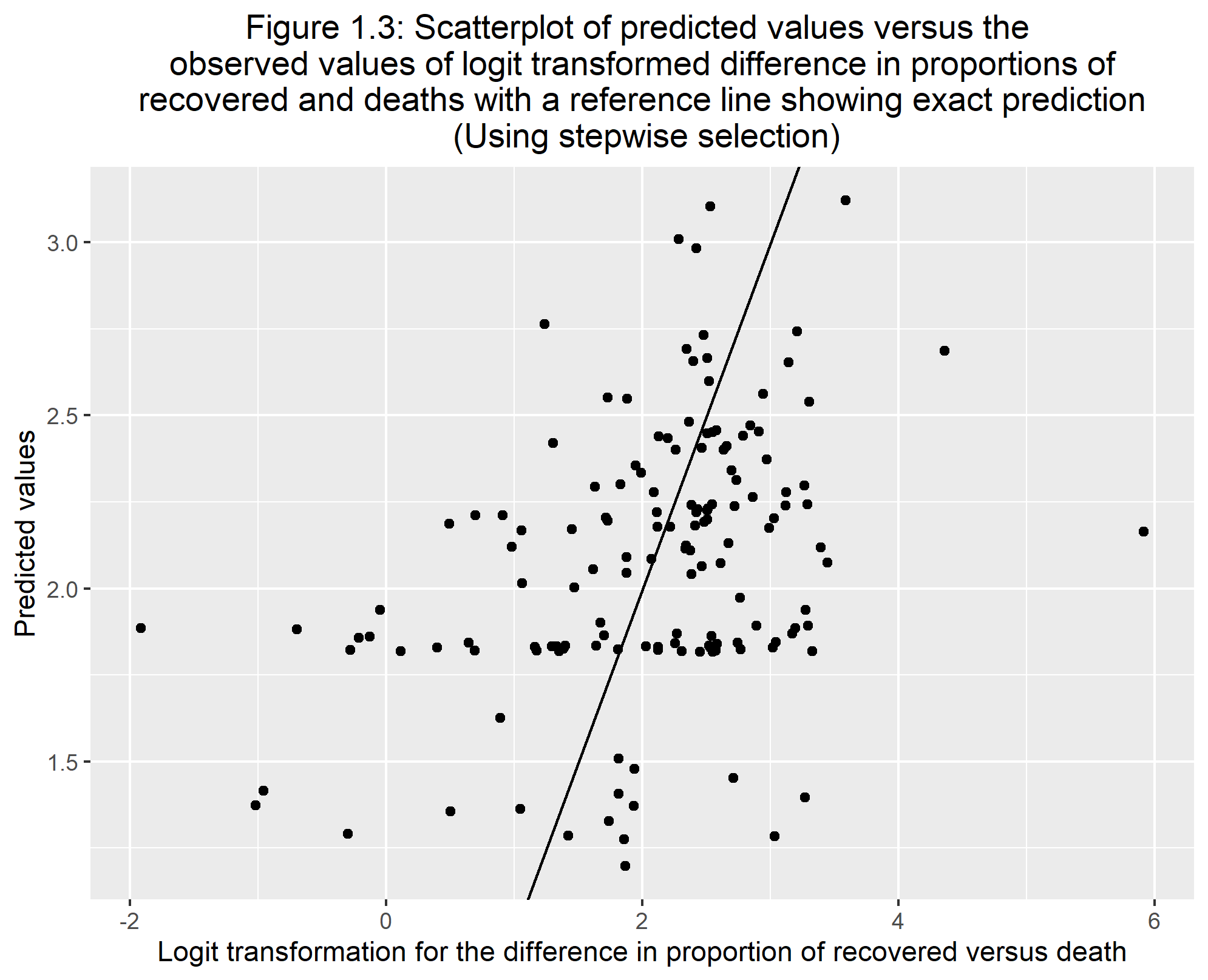
Table 1.3 shows the resulting model from the step-wise selection. If you want more details on this selection, there is a table showing the process of model selection within the supplementary material. We can see that the two variable that were selected using step-wise selection were percent of cases in a country and the location or continent the country is in. Although these two variables were selected and they are significant on at least one level, we can see that the fit according to R-squared is still very low.

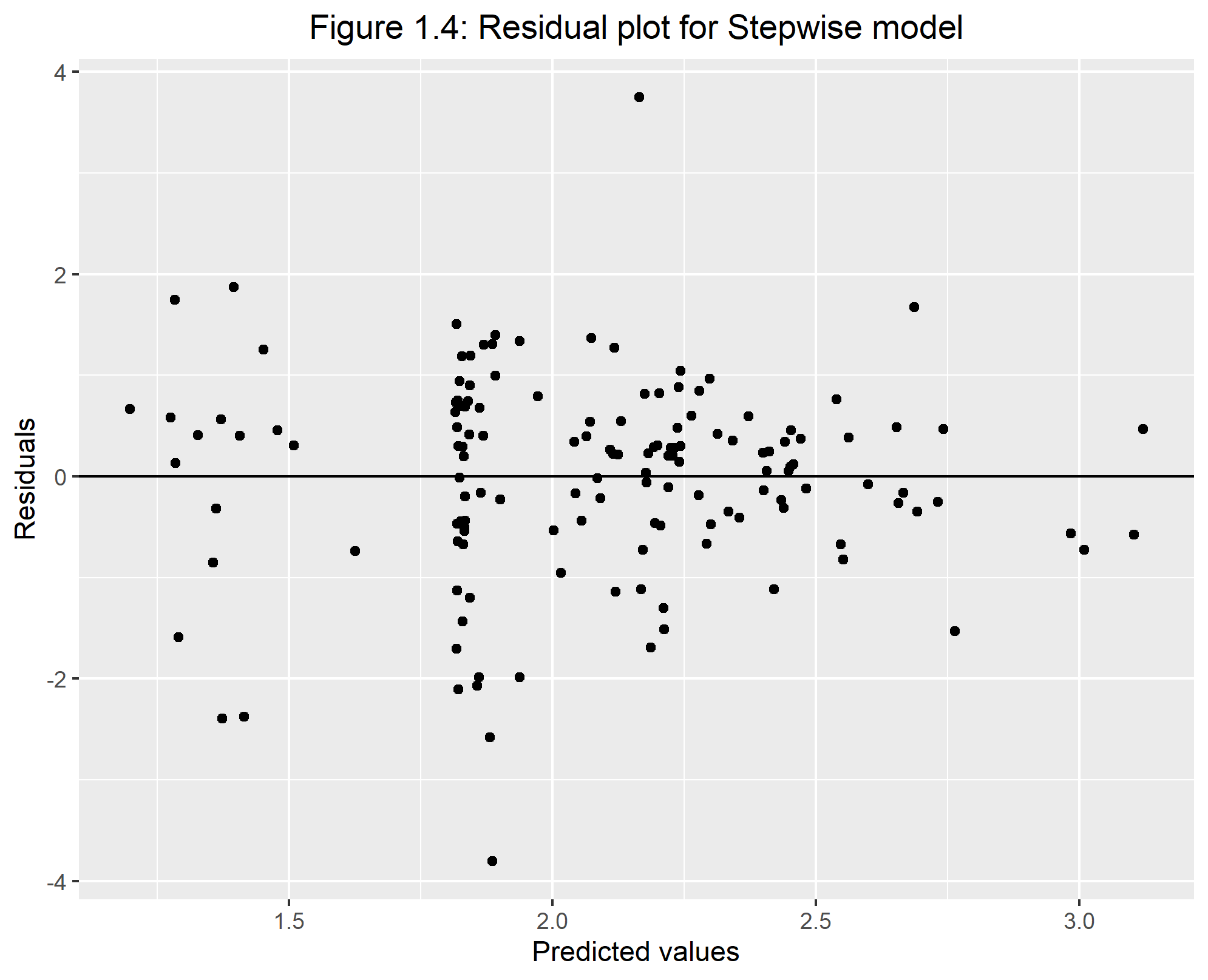
Table 1.3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 1.8145496 | 0.1511012 | 12.0088386 | 0.0000000 |
| pct\_cases | 0.0648311 | 0.0261311 | 2.4809971 | 0.0143033 |
| locationAsia | 0.3283246 | 0.2462683 | 1.3331987 | 0.1846633 |
| locationEurope | -0.0907274 | 0.2970298 | -0.3054489 | 0.7604845 |
| locationN. America | -0.6280091 | 0.3057342 | -2.0541019 | 0.0418534 |
| locationOceania | -0.5438140 | 0.5279834 | -1.0299832 | 0.3048197 |
| locationS. America | 0.0050026 | 0.3879204 | 0.0128959 | 0.9897294 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.1336744 | 0.0960081 | 1.013578 | 3.54891 | 0.0026752 | 6 | -204.1143 | 424.2287 | 448.0425 | 141.7729 | 138 | 145 |

This low fit can be shown within Figure 1.3 and Figure 1.4 where they show the predicted values versus observed values as well as the corresponding residual plot. The data points are spread out and it seems, within the residual plot, it shows a funnel shaped plot which means heteroskedasticity. Therefore, possibly more work could be done to this data/question to get a better fit such as a transformation on the outcome. However, the outcome is already transformed with a logit transformation.





## 3.3 Question 2: Percent of people vaccinated for each country

### 3.3.1 Methods

For my analysis on question 2, I will be using the logit transformation of the proportion of people vaccinated within each country as my main outcome. Since I am using the logit transformation, I am creating another data set that is a subset of the overall data set which only includes the variables I will be using for my analysis. The predictor variables I will be testing out on the main outcome are: percent of cases within a population, continent the country is on, number of tests per person, gdp per capita, proportion of those who died, government type, and health care funds. These will create my full model.

The first part of my main analysis will be using test/training data and finding the RMSE for both the full model and the null model which only include a intercept. This is just to see if using a full model is better than the null model.

After comparing the two RMSE’s, I will be going into more detail on model fitting and tuning. I will be applying my full model to a machine learning technique, the LASSO model. LASSO stands for the Least Absolute Shrinkage and Selector Operator. The reason why I am choosing the LASSO model compared other machine learning models such as a simple tree or a random forest is because LASSO is in the middle of simplicity and complexity which is what I describe the simple tree and the random forest, respectively. Using the LASSO model will try to find the best performing model for the data. This uses a tuning process and a 5-fold cross validation. This process will penalize variables if it is not useful in predicting the outcome. This penalty will not remove the variables but instead penalize these variables to values of zero or close to zero where they wont have a large impact on the model. This will help find the best model to predict the logit transformation of the proportion of people vaccinated. I will be applying this lasso model to both my test and training data to make sure the model fits both data sets similarly so show improvement on over-fitting.

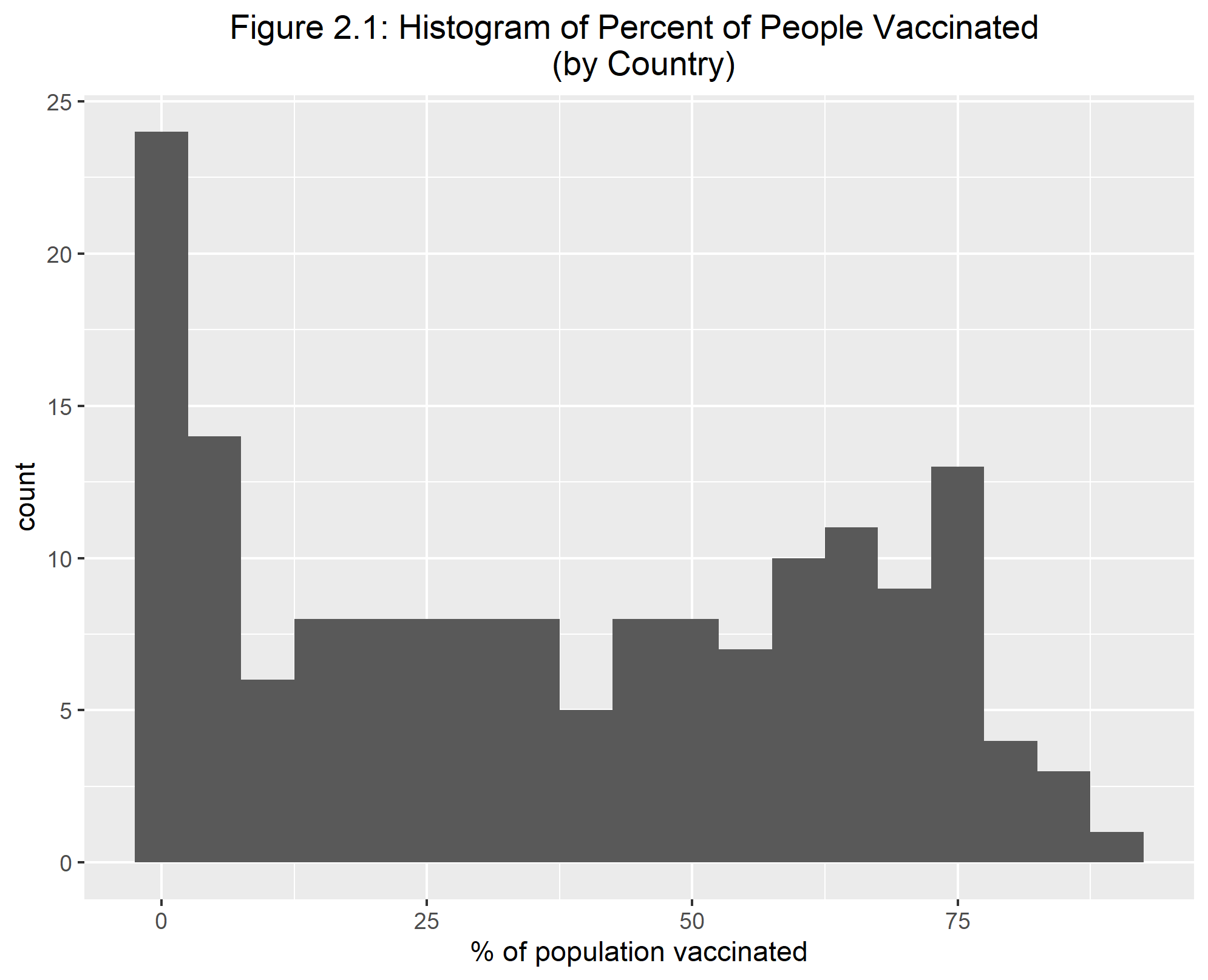
### 3.3.2 Exploratory Analysis

Table 2.1 displays the percentage of people vaccinated within the top five and bottom five countries. Just by the look of the countries displayed, the less developed countries tend to not have a lot of people vaccinated. It is honestly kind of scary to see some countries are not vaccinated at all or very low.

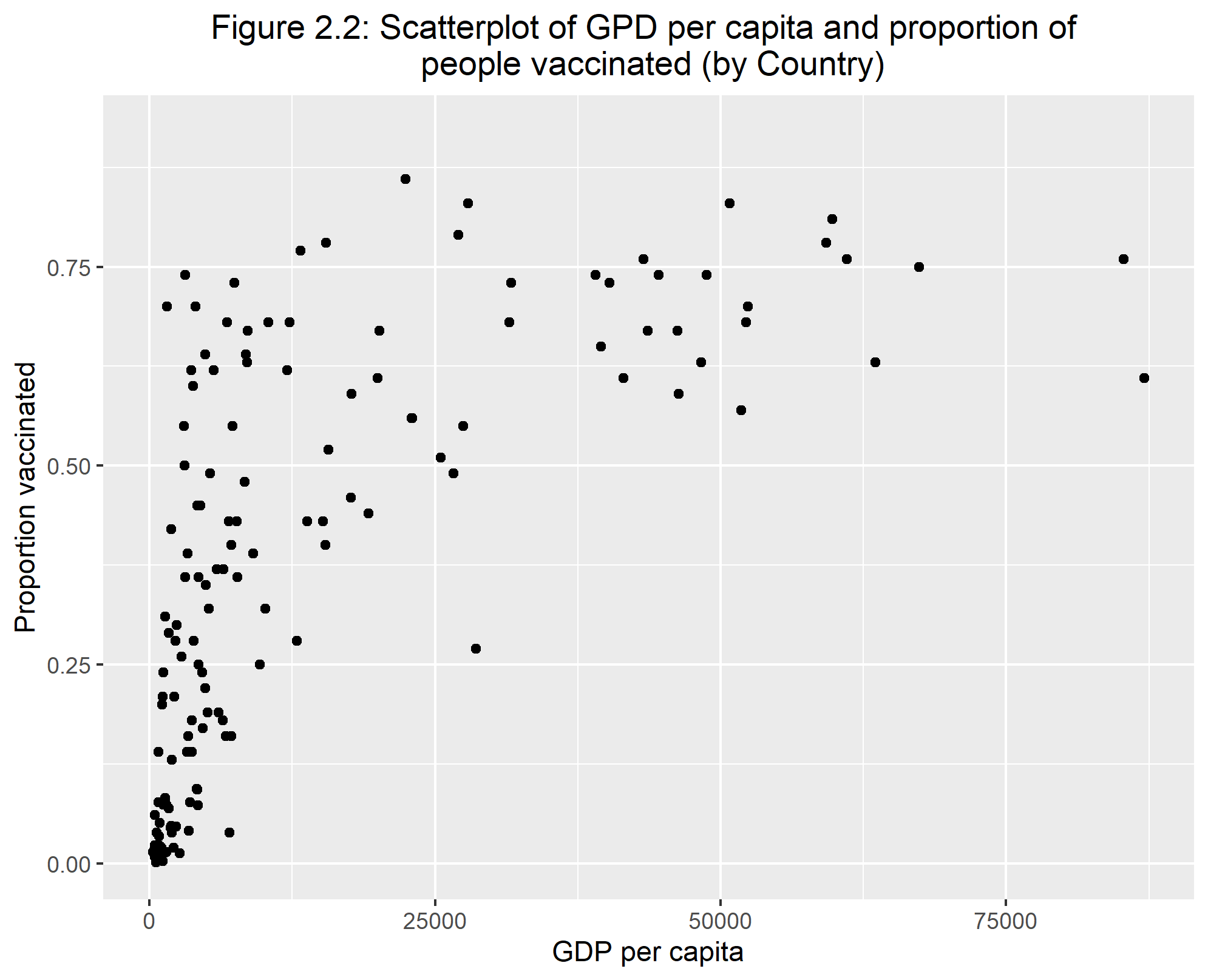
Table 2.1: Top 5 and bottom 5 vaccination percentages by country

|  |  |  |  |
| --- | --- | --- | --- |
|  | Country | location | % of population vaccinated |
| 1 | UAE | Asia | 92.0 |
| 2 | Portugal | Europe | 86.0 |
| 3 | Malta | Europe | 83.0 |
| 4 | Qatar | Asia | 83.0 |
| 5 | Singapore | Asia | 81.0 |
| 159 | Madagascar | Africa | 0.8 |
| 160 | Chad | Africa | 0.5 |
| 161 | South Sudan | Africa | 0.5 |
| 162 | Haiti | N. America | 0.3 |
| 163 | Congo | Africa | 0.1 |

Figure 2.1 shows a histogram figure of the percent of vaccinated people. We can see that the histogram is right skewed with the spike around the percentages of 0%.



When looking at Figure 2.2, we can see that those countries who have large values of GDP per capita tend to have their population more vaccinated. The vaccination proportion varies when the GDP per capita is relatively low.



Looking at Figure 2.3, this side-by-side boxplots show that Africa is different to all of the other continents. Africa is historically known as a less developed continent. This could be result in lower supplies to Africa compared to other places in the world. Within the other continents, Europe seems to have th highest median vaccinated proportions.

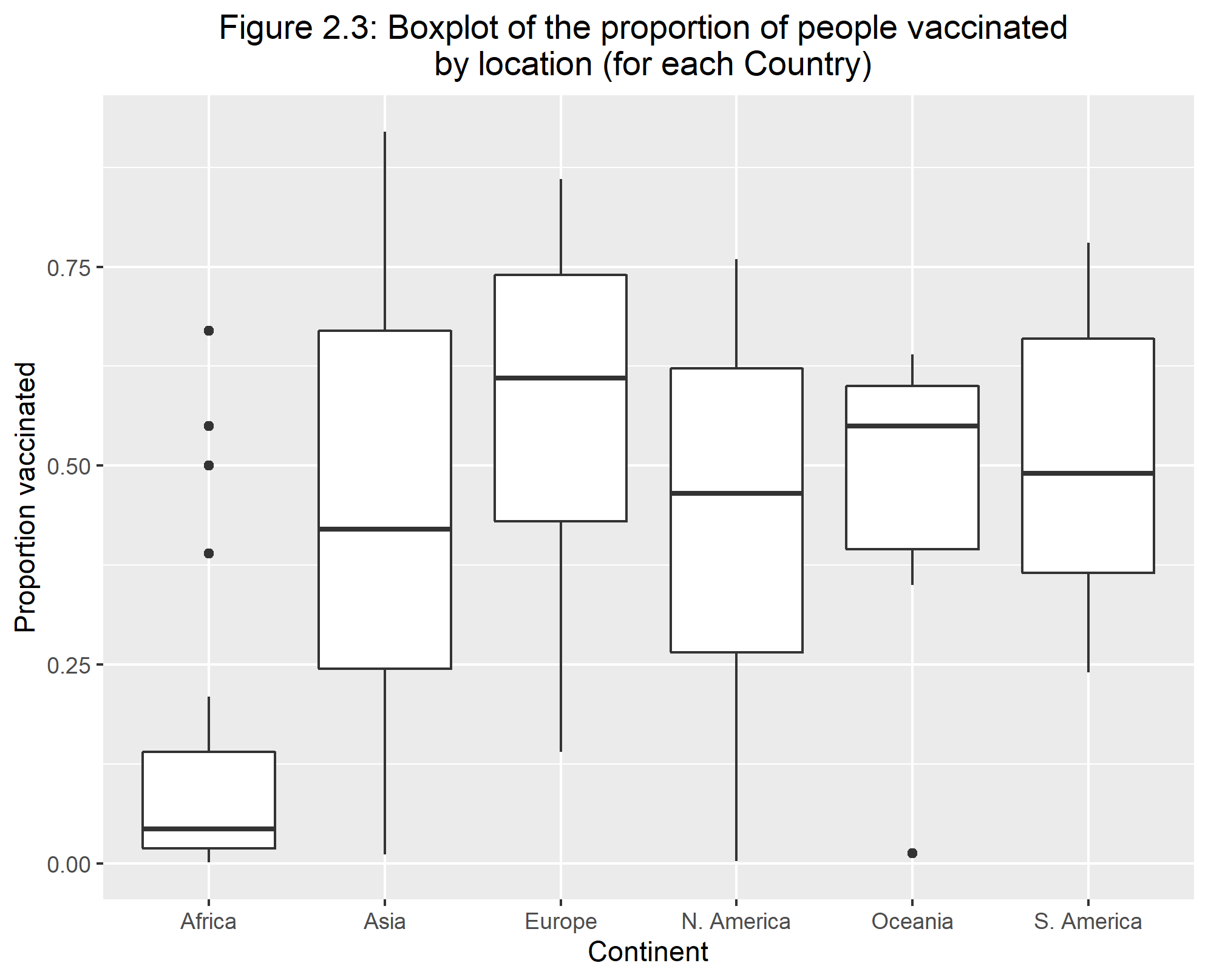
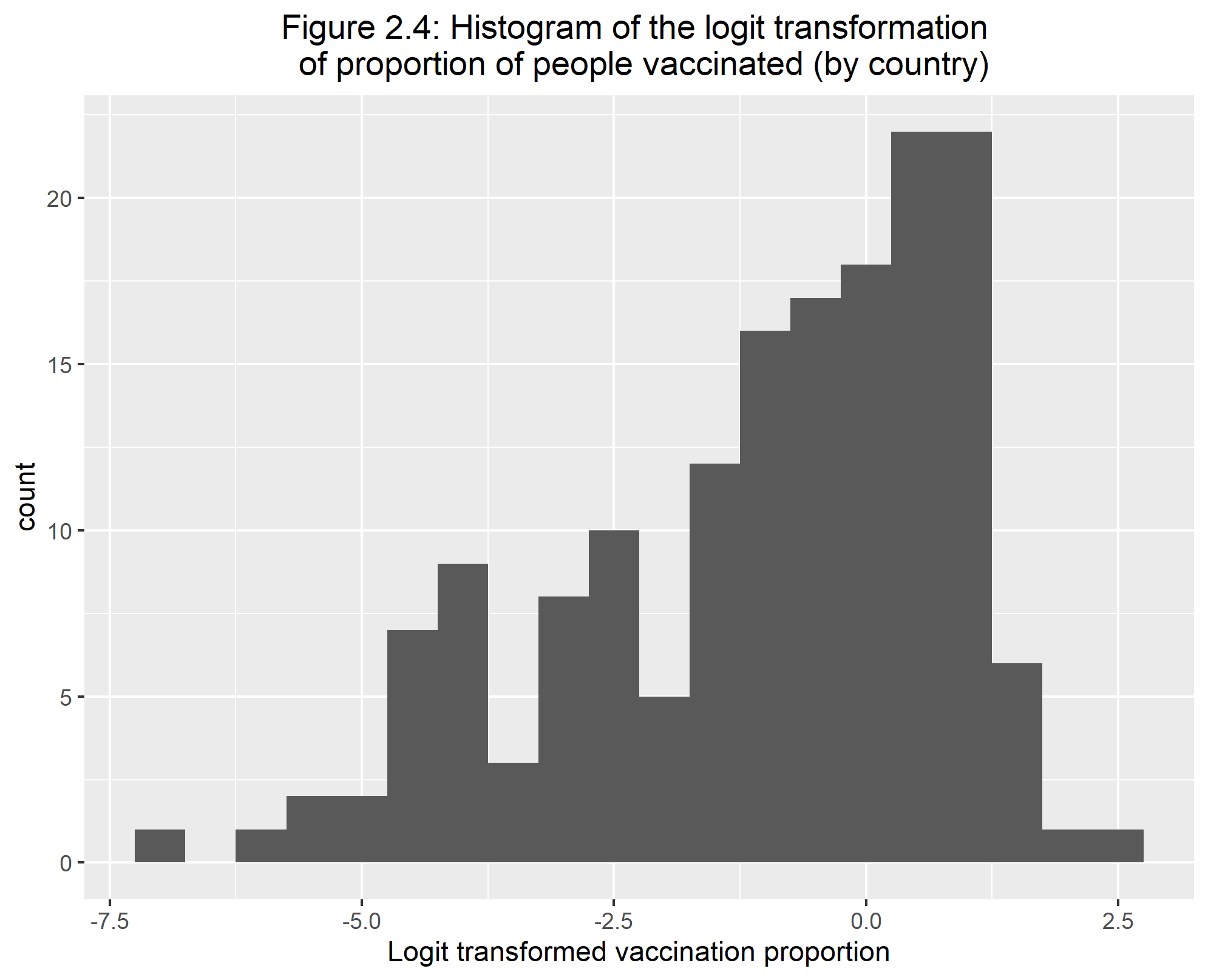


Figure 2.4 is showing the logit transformation of the proportion of the population vaccinated. This is due to the skewness of the non-transformed vaccination numbers. We see that this is still skewed but its skewed left where most of the data is around that 0 value.



### 3.3.3 Main Analysis

As specified within the methods section of question 2, the first part of my analysis will be comparing the RMSE of my full model to the null model. However, before I do this I will run a correlation matrix of the numeric variables to see if some variables are highly correlated.

In Table 2.2 we can see that the two variable, which I suspected, that are highly correlated are health care funds (Value) and GDP per capita. Since those two variables are highly correlated and the health care funds variable only includes 48 observations, I will remove health care funds out of the full model.

Now lets see how our outcome, prop\_vacc, is correlated with our predictors. Proportion of vaccinated people has roughly an 0.57 correlation with both Value which is health care funds and gdp per capita. This was something I was curious about, if countries with a lot of funding could have a direct association to vaccinations. We will see in our main analysis to see if gdp per capita is significant in predicting the proportion of vaccinated people.

Table 2.2: Correlation matrix of predictors and outcome

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | prop\_vacc | pct\_cases | test\_per\_person | Value | gdp\_per\_capita |
| prop\_vacc | 1.0000000 | 0.1684084 | 0.2318597 | 0.5676869 | 0.5756738 |
| pct\_cases | 0.1684084 | 1.0000000 | 0.1536646 | 0.1593376 | 0.0140531 |
| test\_per\_person | 0.2318597 | 0.1536646 | 1.0000000 | 0.2479436 | 0.2898869 |
| Value | 0.5676869 | 0.1593376 | 0.2479436 | 1.0000000 | 0.8735441 |
| gdp\_per\_capita | 0.5756738 | 0.0140531 | 0.2898869 | 0.8735441 | 1.0000000 |

Now that I have my full model solidified, Table 2.3 shows the summary table and some other statistics of this full model where the outcome is the logit transformation for the proportion of vaccinated people.

We can see that our adjusted R-sqaured value is around 0.64 and most of our factors are significant except the proportion of this who died and the COVID tests per person.

Table 2.3: Summary statistics for full model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -2.8714980 | 0.2641891 | -10.8691000 | 0.0000000 |
| pct\_cases | 0.0989961 | 0.0302531 | 3.2722585 | 0.0013594 |
| locationAsia | 1.7113258 | 0.2846914 | 6.0111607 | 0.0000000 |
| locationEurope | 1.3885657 | 0.3774886 | 3.6784307 | 0.0003397 |
| locationN. America | 1.4184950 | 0.3623432 | 3.9147833 | 0.0001438 |
| locationOceania | 0.9185536 | 0.6221918 | 1.4763190 | 0.1422220 |
| locationS. America | 2.3587484 | 0.4477937 | 5.2674885 | 0.0000005 |
| test\_per\_person | 0.0110275 | 0.0698538 | 0.1578660 | 0.8748018 |
| gdp\_per\_capita | 0.0000304 | 0.0000071 | 4.2899945 | 0.0000341 |
| prop\_death | -7.6785506 | 8.6071937 | -0.8921085 | 0.3739460 |
| governmentMonarchy | 0.4722383 | 0.2693520 | 1.7532383 | 0.0818650 |
| governmentProvisional | -0.9040176 | 0.4699679 | -1.9235729 | 0.0565448 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.668835 | 0.6414454 | 1.125072 | 24.41932 | 0 | 11 | -216.571 | 459.142 | 497.8395 | 168.3496 | 133 | 145 |

**Simple Comparison of RMSE**

To start comparing the full model to the null model in terms of RMSE, we need to create training and test data sets. I used a 0.75 split for the training set. I am working on this project using Tidyverse and Tidymodels packages which uses the functions of recipes and workflows. The recipe I am specifying is the creation of the model which in this case, I need to make two. One of the recipes is for the full model and the other is for the null model. Since my outcome is numeric, we need to specify the type of model as well which is linear regression for our project. Once we have our model type and our recipe, we can combine them and put it in a workflow. This workflow will be able to combine the two ‘settings’ we have and create a fit object which eventually creates the statistics of our models and our predictors. However, since our main objective is looking at the fit of our data, we want to look at the RMSE.

We can fit our fit object with both our test and training sets to see how well it does for both data sets. A general consensus on RMSE is that models with a value close to 1 is good and those who go farther away in the positive direction have a worse fit. We want our test and train values to be similar as well as that shows the model does not overfit the training data and results in a lower fit on the test data.

Table 2.4 shows the two values for the full model where the top is the training value and the bottom is the testing value. We can see that these two values are somewhat similar and possibly we want it to be a bit closer. However, these two values are relatively low which is a good sign.

Table 2.5 has the same setup as Table 2.4 where the top value is the training rmse and the bottom is the testing rmse. We can see that these two values are very high but they are somewhat close to each other but not as close as the full model rmse’s.

Comparing the two models, we can see our full model does a lot better than the null model. The RMSE is a lot worse for the null model. This is a good sign for the data as we now know that these predictors do have some type of association since it is way better than the null model. We can now proceed to creating our LASSO model and tuning it to find the best performing model.

Table 2.4: RMSE values for full model (training and test data, respectively)

|  |  |  |
| --- | --- | --- |
| .metric | .estimator | .estimate |
| rmse | standard | 1.080064 |
| rmse | standard | 1.158282 |

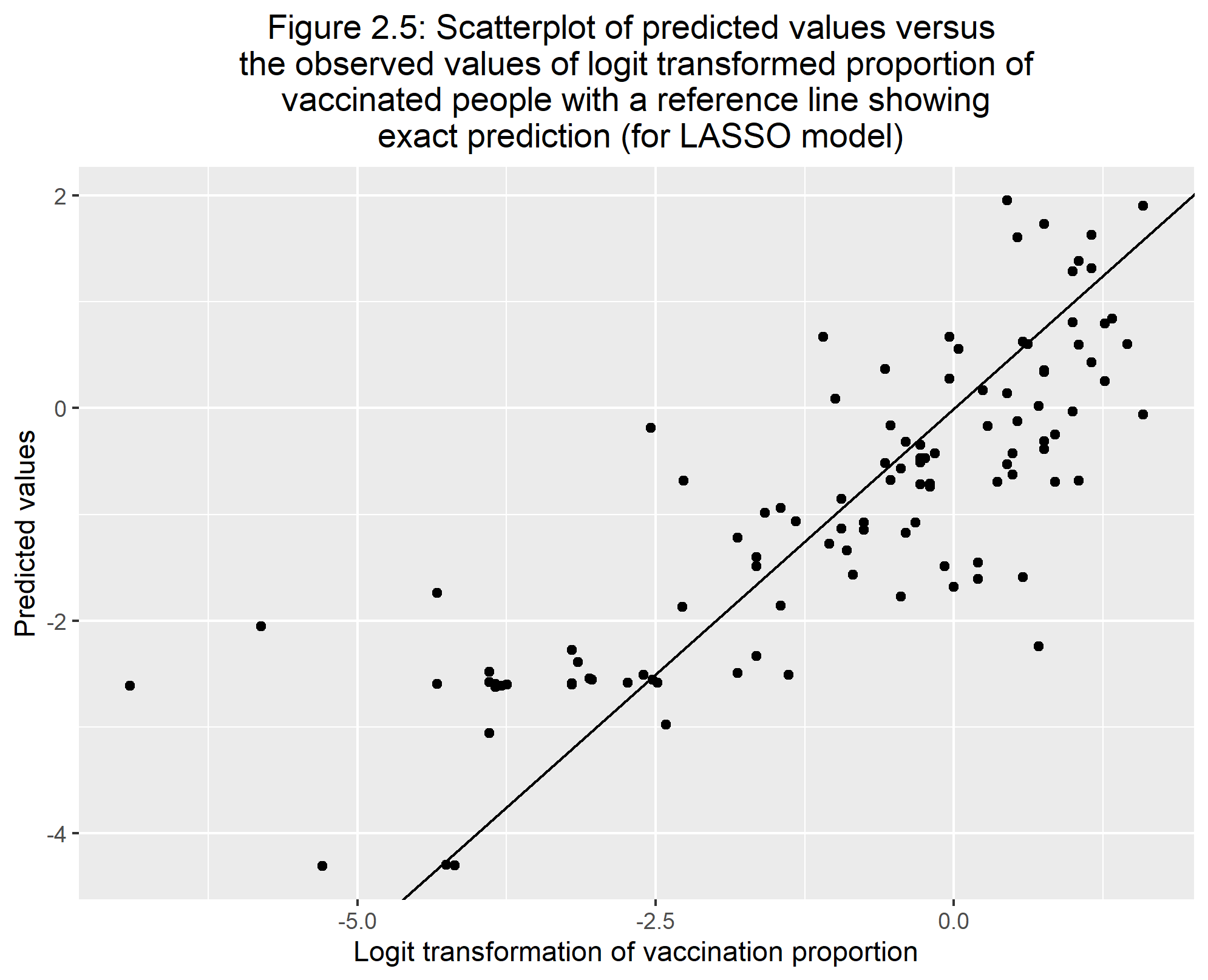
Table 2.5: RMSE values for null model (training and test data, respectively)

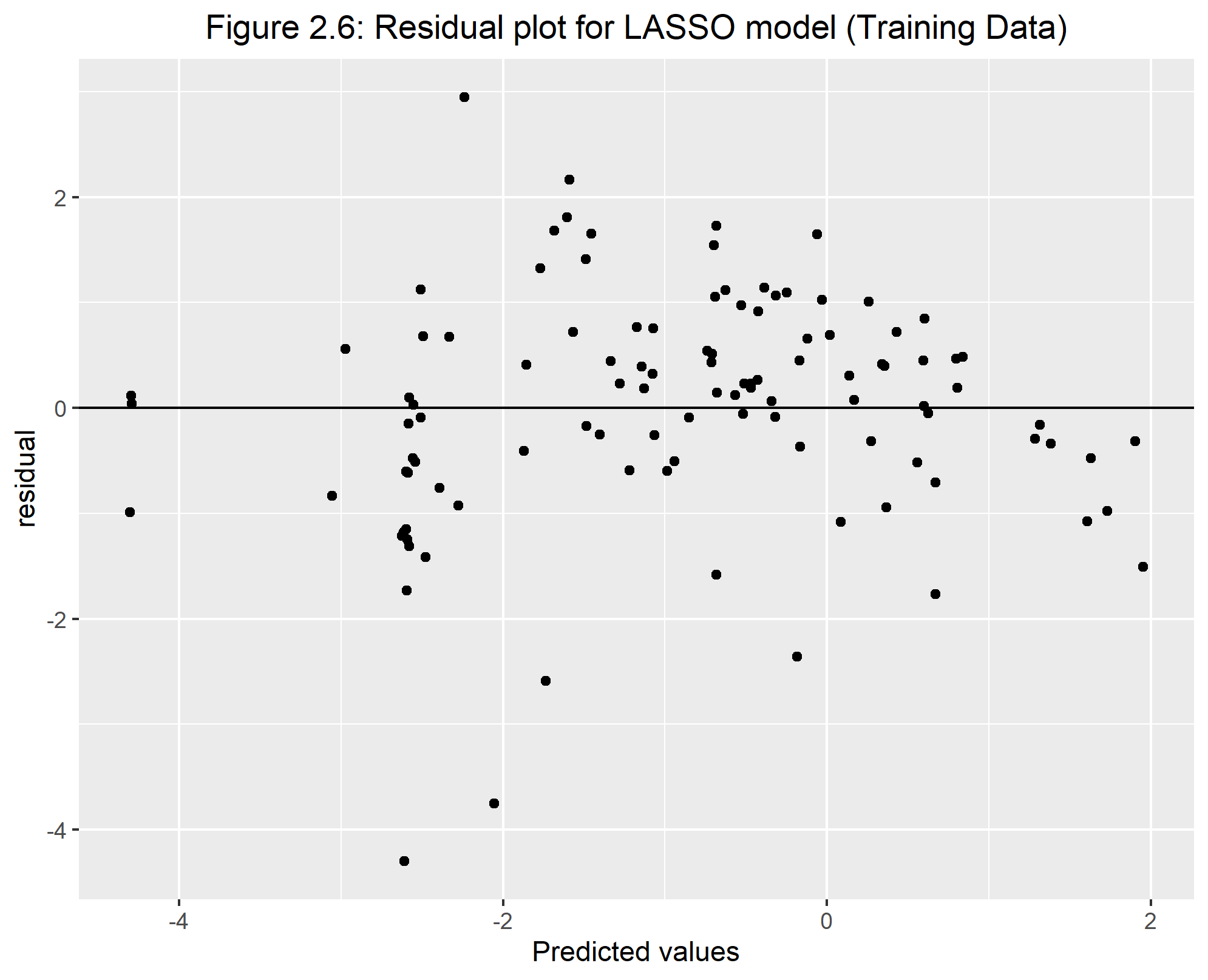
|  |  |  |
| --- | --- | --- |
| .metric | .estimator | .estimate |
| rmse | standard | 1.847835 |
| rmse | standard | 1.954252 |

**LASSO Model**

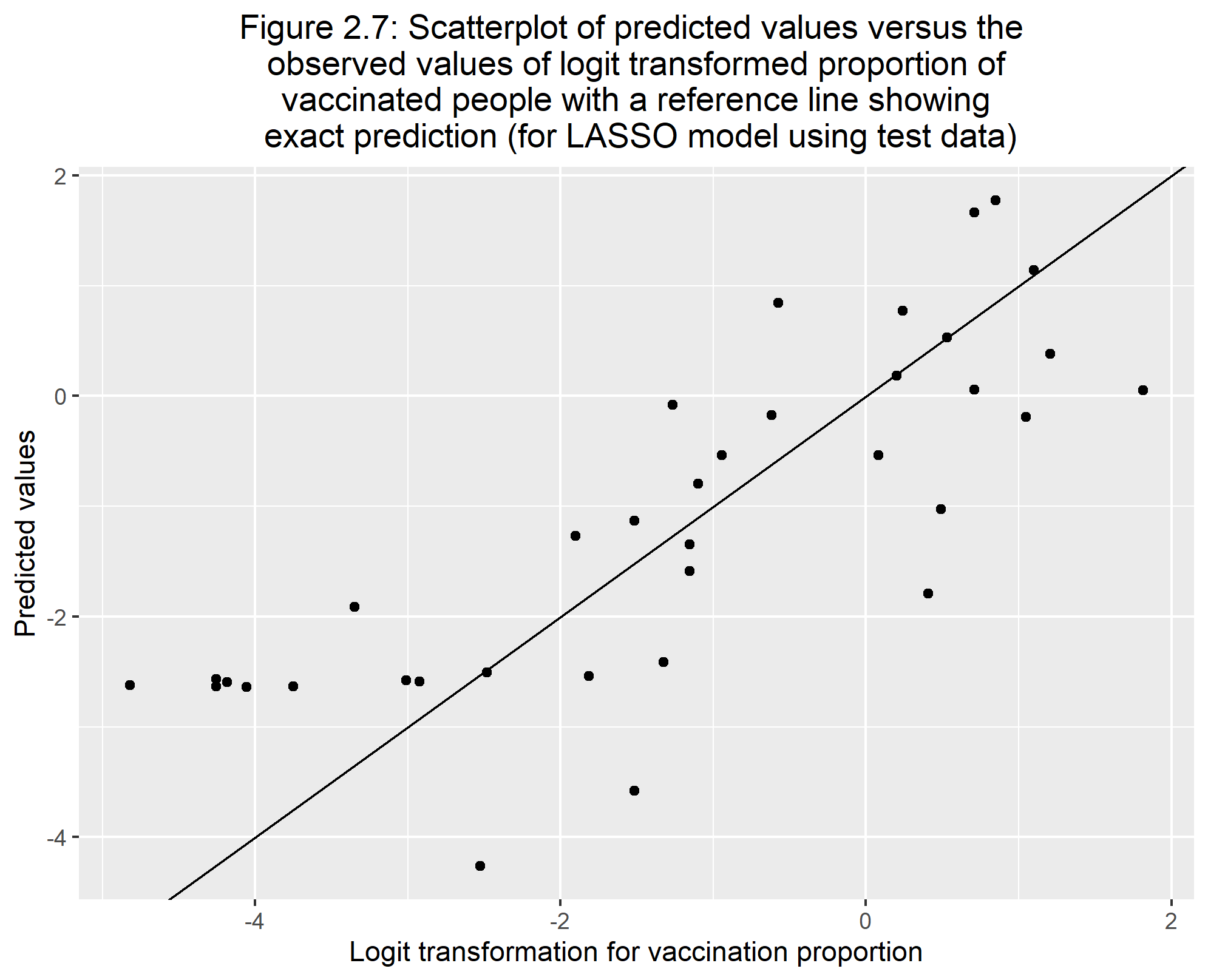
For the LASSO model, there is some preparation we need to run this. We will use cross validation to create the multiple models included in the LASSO model. This cross validation will be a 5-fold stratified by the outcome. We then, like previously, create a workflow using a specific model type and recipe. The recipe will be the full model recipe but the model type will be different. The LASSO model uses a specific model that has a tuning factor into a linear regression type. Once we have our workflow, we need to make a grid as the LASSO model created multiple models based on the 5-fold cross validation and other factors. From the workflow, we can apply a tuning grid to it. This will produce the variety of models. To find the best model, we an apply a function called ‘show\_best’ and specify RMSE to find the best model in terms of RMSE. After putting this model into a workflow and creating a fit object, we get the statistics of the best model.

Figure 2.5 shows the predicted values to the observed values. Most of the points follow the trend and there seems to show no pattern. We can see this more within Figure 2.6 which is the residual plot. Within Figure 2.6, we could see a slight grouping of the points around the predicted value of -2.5.





The last part of the LASSO modeling analysis is fitting the model to the test data to see how well they both look together. To look at the test data results, we need to run a ‘last fit’ function which applies the test data to the best model found above. Figure 2.7 and Figure 2.8 show the predicted values versus observed and its residual plot for the test data. Just like the training data, most of the predicted values do have a similar trend and it close to the observed value. The residual plot shows that grouping of observations for the lower values. This residual plot show more variability in the lower values as well.



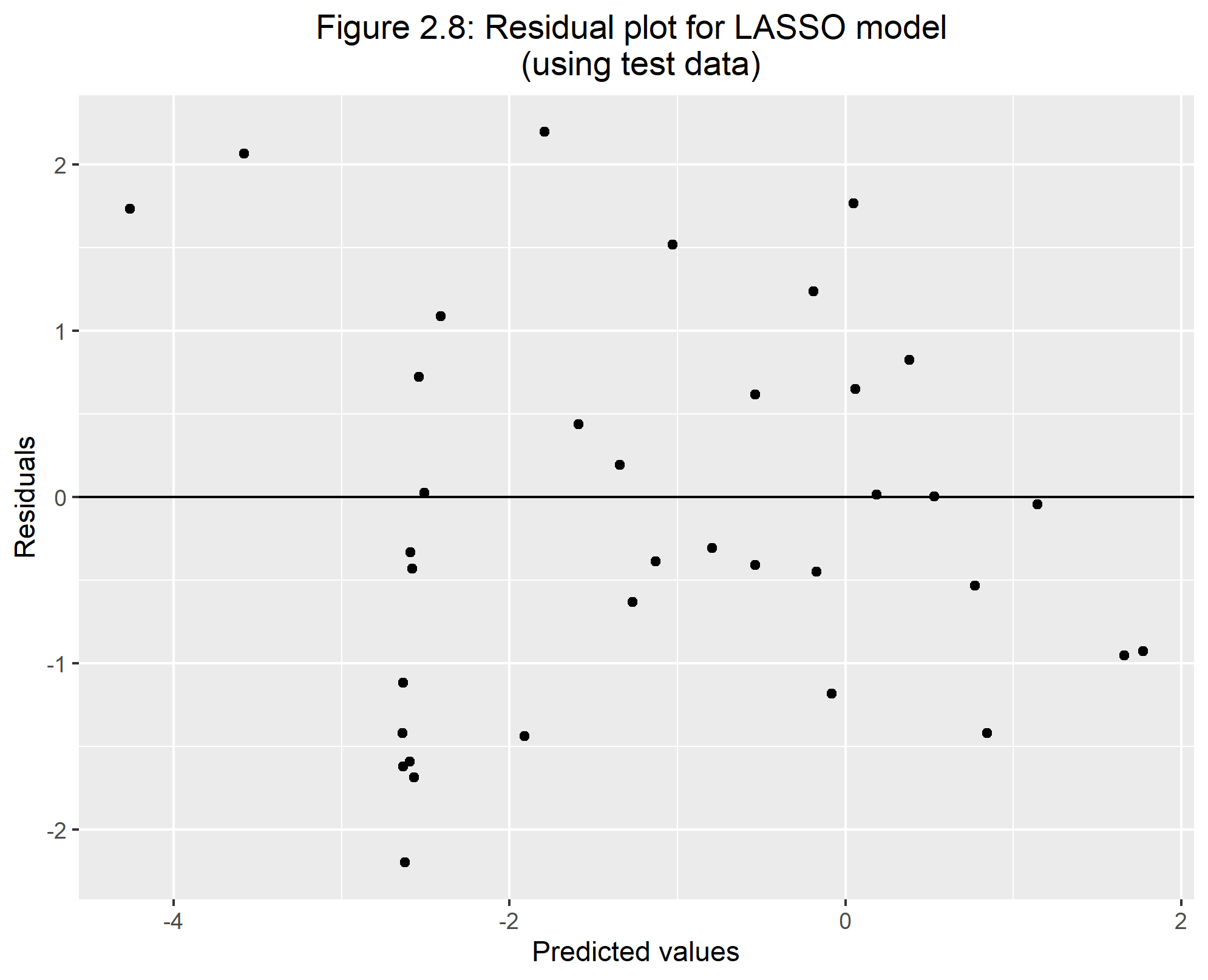


Table 2.6 shows the RMSE values created from this LASSO model. We can see unfortunately these are overall higher compared to the full model numbers which were around 1.08 and 1.16 for the training and test data, respectively. It is still better than the null model however (1.85 and 1.95).

Table 2.6: Comparing RMSE for training and test split data for LASSO model, respectively

|  |  |  |  |
| --- | --- | --- | --- |
| .metric | .estimator | .estimate | .config |
| rmse | standard | 1.218491 | Preprocessor1\_Model27 |
| rmse | standard | 1.187789 | Preprocessor1\_Model1 |

This is the product from the LASSO modeling. This is the model produced by LASSO. In Table 2.7, we can see there are 4 predictors and 2 were removed. The predictors that were removed was the proportion of those who died in each country and test per person. However these estimates are based on the logit transformation of our outcome. The last column is the transformation from the logit value to help in more practical predictions.

From the estimates, we can see although there was 2 variables that suppressed to a value of 0 based on the LASSO model, the predictor gdp per capita have roughly a value of 0. Therefore the LASSO model did surpress this variable as well. It just was not exactly to 0. We can see the continent that the country is in also has an affect on proportion vaccinated where the continent Africa is the reference group. The continents that increase the odds for a higher proportion of being vaccinated the most are Asia and South America. We can also see that in terms of government type where a republic government is the reference group, those countries that are a monarchy increase the odds of a higher proportion of vaccinations. In opposite affect, a provisional government country lowers the odds of getting more people vaccinated proportionally.

Table 2.7: Summary statistics of LASSO model and predictors with their respected transformed values

|  |  |  |  |
| --- | --- | --- | --- |
| term | estimate | penalty | Transformed\_value |
| (Intercept) | -2.6588215 | 0.048939 | 0.0700307 |
| pct\_cases | 0.1342715 | 0.048939 | 1.1437033 |
| gdp\_per\_capita | 0.0000316 | 0.048939 | 1.0000316 |
| location\_Asia | 1.2032127 | 0.048939 | 3.3308006 |
| location\_Europe | 0.5995112 | 0.048939 | 1.8212284 |
| location\_N..America | 0.5420638 | 0.048939 | 1.7195520 |
| location\_Oceania | 0.1796438 | 0.048939 | 1.1967910 |
| location\_S..America | 1.4786906 | 0.048939 | 4.3871972 |
| government\_Monarchy | 0.6308822 | 0.048939 | 1.8792677 |
| government\_Provisional | -1.6703589 | 0.048939 | 0.1881795 |

# 4 Discussion

## 4.1 Summary and Interpretation

The analysis was split up into two different analyses specified by the two questions.

The first analysis was looking at the difference of proportions between recovery and death rates. With some exploratory analysis, there was low fitting according to the simple linear regression between multiple predictors and the outcome. Due to the very small r-squared values, I decided to do a step-wise selection on the predictors. In result, I found that percent of cases and the continent is the remaining two variables. The only level from the location variable that is significantly different from the reference group, Africa, is North America. Although they show a significant impact in predicting the logit transformation of difference in proportions, we can see that step-wise selection does not create a model that fits the data well. While looking at the residual plot, we see a funnel shaped plot which means heteroscedasticity and possibly more transformation options could be applied to make this model fit better. However in this current state, the variables do not fit the data well. If we look at the estimates and transform them out of the logit transformation, we find that if we increase the percentage of cases within a country, there will be a slight increase in odds of having a higher recovery proportion compared to the death proportion with the location staying constant. Contrary, North America countries, compared to African countries as a whole, decreases the odds of having a higher difference in proportions between recovery and death.

The second question pertains to how vaccination proportions are affected by multiple predictors. Using the logit transformation and applying the LASSO modeling, out of the 6 predictors within the full model, only 4 remained within the equation. These include percent of cases for a country, gdp per capita, the continent the country resides, and the type of government. Out of these, those who have the most impact are location and government type, and gdp per capita. However, gdp per capita has roughly an estimate of 0 which probably has no impact on the proportion of people vaccinated. With the reference for the location variable being Africa, we can see that no matter which level of location you look at, they increase the odds of having a higher proportion of people getting vaccinated while keeping all other variables constant. The type of government being significantly different could be due to just the small amount of countries that classify as monarchy and provisional compared to republic or it could really be that significantly different. Regardless, the monarchy type of government increase the odds that the proportion of vaccinated people is higher compared to a republic type of government, and a provisional government decreases the odds that the proportion of vaccinated people is higher compared to the republic government given all other variables constant.

## 4.2 Strengths and Limitations

Some strengths to this analysis was that since this topic of COVID-19 is in today’s news, a lot of each country’s information is available in public use. Along with COIVD-19 information, most information of countries are of public use. Therefore a lot of information was easy to find.

Some limitations for this analysis was that I was using cumulative data which means as more days pass, this analysis will become old. Another limitation was although I got the information I wanted, I had to use multiple sources to find them which results in combining data sets and ultimately errors.

## 4.3 Conclusions

We found that the percent of cases and the continent are significantly associated with the difference in proportion of recovered and deaths. The higher the percent of cases there is in a country, the higher recovery proportion there is when keeping the continent constant. North American countries as a whole has a higher death proportion compared to African countries as a whole when keeping the percent of cases constant. I was curious on if the amount of health care support helps patients recover. However based on these results, we see that the amount of tests per person and gdp per capita does not really affect recovery rates unfortunately. However, the fit from my resulting step-wise selected model is still not good. This is constant throughout the simple linear regressions and the step-wise model. A different transformation could possibly help better the fit more as the residual plot shows a funnel shaped plot.

If we revisit the Liang et al. paper, they mentioned in their analysis that the tests per person, type of government, and economic status negatively impacts mortality rates (Liang, Tseng, Ho, & Wu, 2020). We see that within my analysis of the difference between recovery and death proportions, none of these variables had an impact on the this difference in proportions. This is not a direct comparison between analysis, and their analysis was during 2020.

We see that the proportion of people vaccinated is associated with the amount of COVID-19 cases, which continent the country is in, and type of government. Given all other variable constant, Asia and South America gives the highest proportional increase compared to Africa. Compare to what Mathieu et al. said, as I previously mentioned above in the general background section, we do see that Africa has a significant different proportion of people vaccinated and in the negative direction. Monarchy governments increase vaccinations compared to republic government countries but provisional governments decrease the proportion of those vaccinated. One prediction I had within the questions/hypotheses section above was that countries with more gdp per capita will have higher vaccinated proportions. However, based on my LASSO model, gdp per captia was not significant in predicting vaccine proportions. If we look at more previous studies I mentioned above in the background section, the United Nations mentioned that the distribution of vaccines could be based on the actual prices of the vaccines. This could have a direct correlation to the gdp per capita for each country. However within my analysis, gdp per capita does not have much of an impact on vaccination proportions when all other variables stay constant (“COVID vaccines,” 2021).

# References

Choi, E. M. (2021). COVID-19 vaccines for low- and middle-income countries. *Trans R Soc Trop Med Hyg*, *115*(5), 447–456. <https://doi.org/10.1093/trstmh/trab045>

COVID vaccines: Widening inequality and millions vulnerable. (2021). https://news.un.org/en/story/2021/09/1100192.

Liang, L.-L., Tseng, C.-H., Ho, H. J., & Wu, C.-Y. (2020). Covid-19 mortality is negatively associated with test number and government effectiveness. *Sci Rep*, *10*(1), 12567. <https://doi.org/10.1038/s41598-020-68862-x>

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., … Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nat Hum Behav*, *5*(7), 947–953. <https://doi.org/10.1038/s41562-021-01122-8>