Factors Affecting COVID-19 Cases Proportion in a Country

Qi Wang[1]
Department of Statistics, University of California, Santa Cruz[1]

Abstract

In this report, we are using a Bayesian regression approach to explore what is the most significant factors affecting the logarithm of the proportion of covid cases in 38 countries. Since there are tons of covariates, to begin with, a Lasso regression is carried out for variable selection. After selecting the variables, both random effect model considering the continent factor and common Bayesian regression model excluding the continent factor are considered. By WAIC, elppd and DIC, the random effect model performs better. Furthermore, it seems United States is an outlier for this dataset, after deleting the United States, with random effect model, it predicts better with smaller RSS than deleting another country, Sri Lanka. Finally, we got the conclusion that two most significant factors are hospital beds proportion and humanity development index, which has a positive correlation with the proportion of covid cases in each country.

KEY WORDS: Random Effect Model, Bayesian Model Selection, Lasso Regression, Hierarchical Model

1.Background and Data Overview

We are going to study the relationship between proportion of total covid cases and other indexes among 38 different countries. The dataset contains information about COVID-19 case, death, testing,and vaccination information as of 1/1/2022 for 38 countries. We want to explore any association between the proportion of total cases and the country-specific covariates. Our complete COVID-19 dataset is a collection of the COVID-19 data maintained by Our World in Data. Furthermore, part of the data also comes from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). We have 38 data and 26 variables(including the response variable) in total. Most of the variables describe one relevant index about covid, like the vaccine coverage rate, total case of covid and so on. Here, we take the proportion of cases in each country as response variable, and others to be the possible covariates to explore the inner relationship between them in a Bayesian regression way.

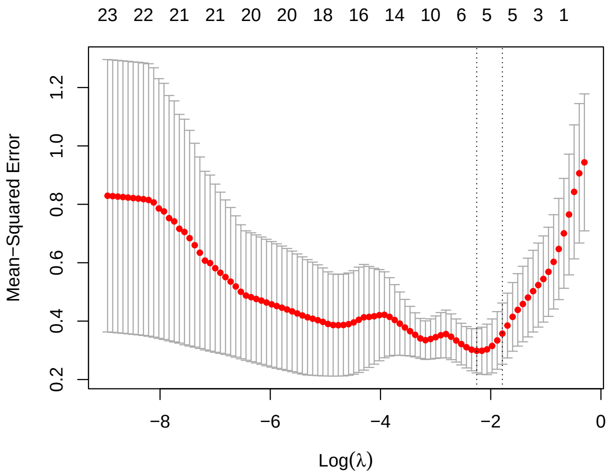

Figure 1: Lasso Regression MSE Trend with Lambda

Table 1: Lasso Regression Result

(Intercept)	-6.677
tests__per__case	-0.002
total__boosters__per__hundred	0.016
male__smokers	0.006
hospital__beds__per__thousand	0.113
human__development__index	3.855

2.Exploratory Data Analysis

Since there are too many covariates comparing with the sample size, before fitting the Bayesian regression, a variable selection process is necessary. I will use Lasso regression to decide which variables will be included later in the Bayesian part. To get the best penalizing constant, a cross validation process is considered. Furthermore, the relationship between MSE and the penalizing constant and the coefficient trace plot in Lasso and ridge regression are in figure 1, 2. After checking the variables selected by Lasso model, as in 1 we can see the humanity development, positive rate, hospital beds proportion diabetes prevalence and people who aged over 70 are five most significant variables both in Lasso model and ridge model. Therefore, I am going to use these five variables in the following Bayesian part.

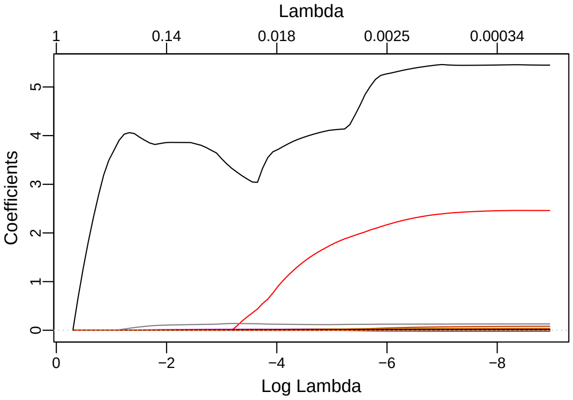

Figure 2: 5 Variables Selected by Lasso

3. Bayesian Regression

3.1 Ignorning Continent Factor Case

3.1.1 Model Setting

In this way, we simply ignore the continent affect, therefore, we set a Bayesian regression model with following settings:

$$Y_i \sim^{iid} TruncN(\beta_0 + X_i\beta_{others}), \sigma^2), Y_i < 0$$

However, this model may not be a good choice since the posterior for β would not be in closed form. I will use a simpler one:

$$Y_i \sim^{iid} N(\beta_0 + X_i\beta_{others}), \sigma^2)$$

in which, X_i includes the total cases, tests per case, total boosters per hundred, male smokers, hospital beds per thousand and human development index of the i^{th} country. And we put a prior on all the β and σ^2 as follows:

$$\pi(\beta_0, \beta_{others}, \sigma^2) \propto \frac{1}{\sigma^2}$$

3.1.2 Posterior Distribution

After using the Gibbs sampler, we got the posterior distribution of all the slope, intercept and noise as in figure 3 in the truncated case. The distribution of β seems very strange, and also the σ^2 concentrates around a big value, which means that our model is not that accurate. The not truncated case instead has the posterior distribution in figure 4. From the data we know all y's are concentrated around -2 to -1, according to the posterior of σ^2, by empirical rule, the probability of getting a positive y based on our regression coefficient is small, so we actually don't need the truncated distribution assumption for y. Instead, we use a normal distribution, which can also provide us the conjugacy.

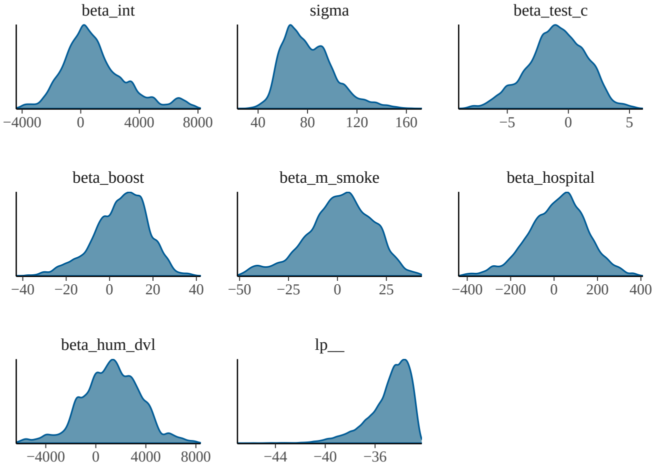

Figure 3: Posterior(Truncated Y and Continent Excluded)

We can see that the tests per case(inverse of positive rate) has a negative effect on the population case proportion, but booster proportions, male smokers, hospital beds proportion and human development index is making the case proportion larger. However, the result seems very counter intuitive. Going back to the data, the continent could also be a very important factor because covid is contagious, so countries from the same continent may have similar trend, which will be discussed in the next section.

3.2 Including Continent Factor Case

3.2.1 Model Setting

In this case, since continent is a categorical model, I am thinking of using a random effect model. First, there is a trend that countries on the same continent will have the similar intercept but not exactly the same intercept among them. We can regard each continent as a group and each country is the group member in the group, so it's reasonable that each group has different intercept. Furthermore, each continent also seems to have their own slope because countries within the same continent tends to work together or take similar policies to fight against the covid. Therefore, our model could be expressed as:

$$Y_{ij} \sim^{iid} N(\beta_{0i} + X_{ij}\beta_{others,i}), \sigma^2)$$

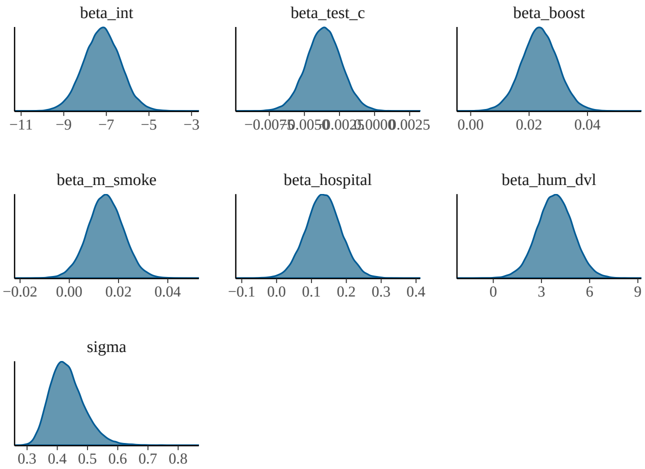

Figure 4: Posterior(Non-truncated Y and Continent Excluded)

$$\beta_{0i} \sim N(\mu_0, \tau_0^2)$$

$$\beta_{ji} \sim N(\mu_j, \tau_j^2)$$

$$\pi(\mu_k) \propto 1,\ \pi(\tau_k) \sim \frac{1}{\tau_k^2},\ k \in \{0, 1, 2, ...\}$$

3.2.2 Posterior Distribution

I simply did not add the interaction term between continent and other variables in my random effect model because the interaction effect between a categorical variable and a continuous variable can be added up together with the fixed effect into the random effect. Therefore, in this setting, I used both random intercept and random slope. The posterior distribution of all μ_i's are as in figure 5. They are tests per case, total boosters per hundred, male smokers, hospital beds per thousand, human development index and intercept from upside down.

4. Model Comparison and Outliers Arrangement

4.1 Model Comparison with DIC, WAIC and elppd

Since we have two models for now, we will need the DIC(Deviance Information Criterion), WAIC(Watanabe-Akaike Information Criterion) and elppd(Expected log pointwise predictive density).

$$DIC = -2log(y|\hat{\theta}) + 2p_{DIC}$$

$$p_{DIC} = 2log(p(y|\hat{\theta})) - \frac{1}{S}\sum_{s=1}^{S} log(p(y|\theta^s))$$

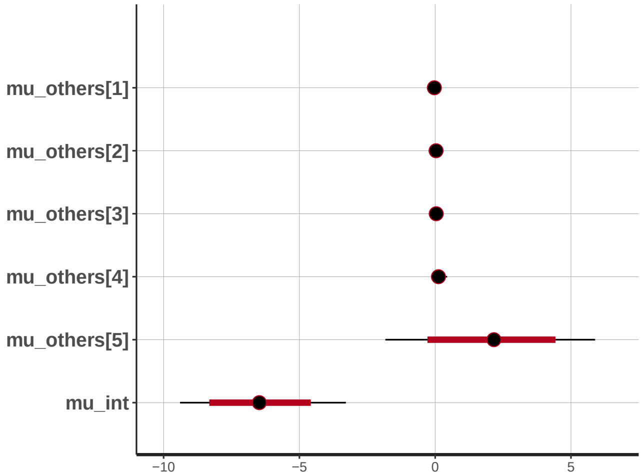

Figure 5: Random Effect Mean Posterior Distribution

Table 2: DIC and WAIC Criteria Model Comparisons

	WAIC	elppd	DIC
Excluding Continent	50.65131	-25.32565	49.51465
Including Continent	38.30105	-19.15052	40.48848

$$WAIC = -2log(p(y|\hat{\theta})) + 2p_{WAIC}$$

$$p_{WAIC} = \sum_{i=1}^{n} var_{\theta|y}(log(p(y_i|\theta^s)))$$

Since we are using a normal likelihood, the mode is just the mean, the distribution of the log likelihood of all iterations are shown in figure 6. After calculating the three criteria, as shown in table 2, we can see that the random effect model including the continent is better in WAIC, elppd and DIC. Therefore, in the later part of regression without outliers, I will use the random effect model,

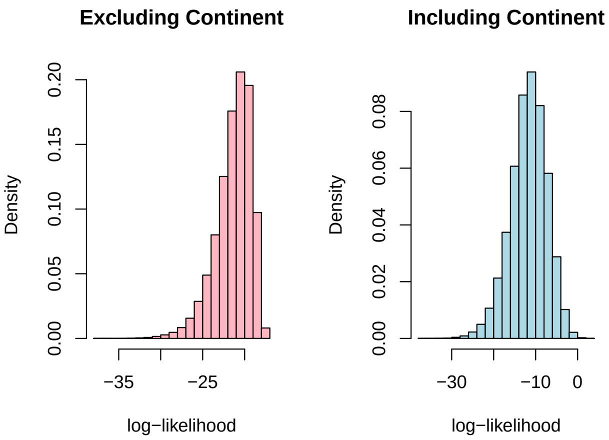

Figure 6: Log Likelihood Distribution of All Iterations

4.2 Random Effect Model Predictive Performance without Certain Countries

4.2.1 Without United States

In this case, we simply delete United States from the dataset since it seems that it is an outlier and may have potential effect on the model. The posterior predictive distribution summary are shown in tables in the appendix for each country except for US. And the overall posterior predictive mean trend compared with the true data is shown in figure 8 and 9. If we take predictive mean to be the predicted value, the residual sum squares as follows:

$$RSS_{-US} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 2.026995$$

Furthermore, the specific prediction for United States is in figure 7. We can see the predictive distribution cannot predict so accurately because US is actually an outlier.

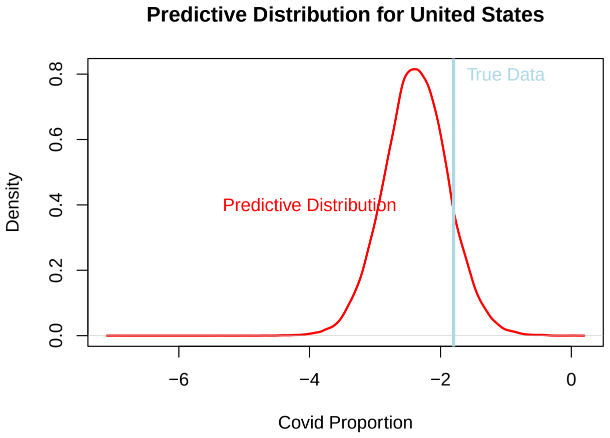

Figure 7: Predictive Distribution for United States

4.2.2 Without Sri Lanka

In this case, as in section 4.2.1, we discard the Sri Lanka and then make a predictive distribution for other countries. The posterior predictive distribution summary are shown in tables in appendix for each country. And the overall posterior predictive performance is shown in figure 11 and 11. Also the RSS of this model could be expressed as:

$$RSS_{-SL} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 2.127073$$

Furthermore, the specific predictive distribution for Sri Lanka is in figure 10

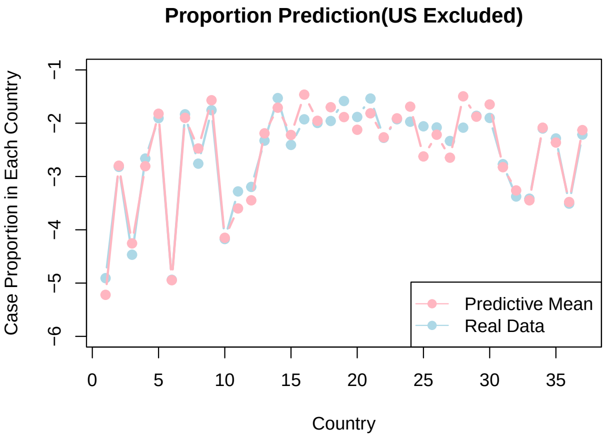

Figure 8: Predictive Distribution(US Excluded)

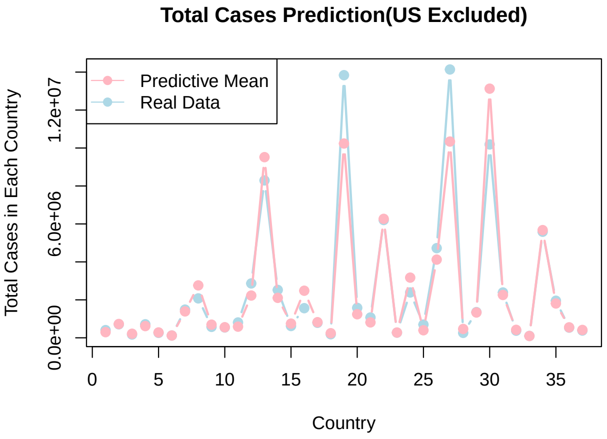

Figure 9: Predictive Distribution(US Excluded)

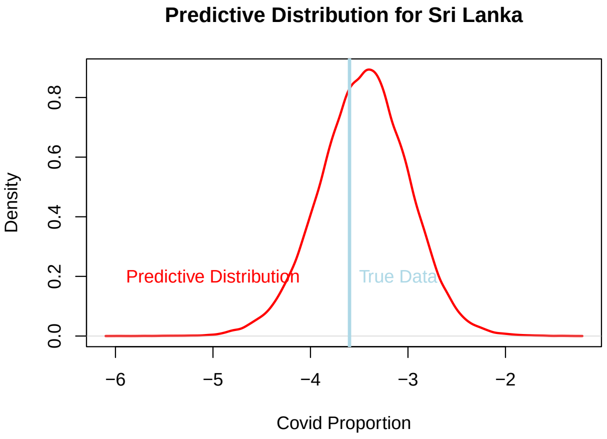

Figure 10: Predictive Distribution for Sri Lanka

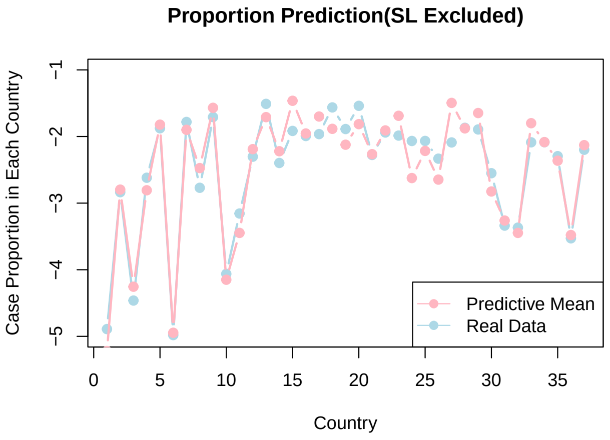

Figure 11: Predictive Proportion(SL Excluded)

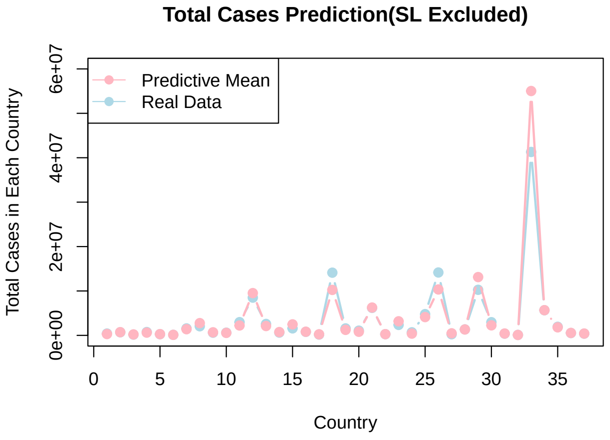

Figure 12: Predictive Total Case(SL Excluded)

5. Conclusion

In all, the most significant coefficient that affects the proportion of the covid is the human development index. This could be interpreted that as the human development index increases, the people are moving from one area to another area of the country more frequently and having more connections to other people, which makes covid more easy to spread. Also, the hospital beds proportion is also positively correlated with the covid proportion, this could be explained by the positive relationship between the testing ability and proportion of beds. As the proportion of beds increases, the ability of doing the test also increases, which means that more covid cases will be found and reported.

Furthermore, as an outlier, United States, if we regress the model without US, the prediction performance will be better. It seems that the model without US will perform better than the model without Sri Lanka because Sri Lanka seems less likely to be an outlier comparing with United States.

6. Discussion

Here I used the random effect model, so I did not consider the interaction effect between continent and others. Imagine this case:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

When X_{1i} is a 0-1 variable, and $x_{1i} = 0$:

$$Y_i = \beta_0 + \beta_2 X_{2i} + \epsilon_i$$

When $x_{1i} = 1$:

$$Y_i = \beta_0 + \beta_1 + \beta_2 X_{2i} + \beta_3 X_{2i} + \epsilon_i = \beta_{int} + \beta_{slope} X_{2i} + \epsilon_i$$

In which,

$$\beta_{int} = \beta_0 + \beta_1, \quad \beta_{slope} = \beta_2 + \beta_2$$

Therefore, we can simply treat this as a random effect model that for each group, it has its own slope and intercept, then the interaction effect has been considered. Also, we even consider the similarity among different groups by fitting a hierarchical model.

0. Appendix

Table 3: Non-US Model Predicting Result

	mean	2.5%	50%	97.5%
Kenya	-4.908	-5.764	-4.913	-4.037
Tunisia	-2.819	-3.752	-2.816	-1.902
Zimbabwe	-4.468	-5.282	-4.468	-3.655
Azerbaijan	-2.663	-3.419	-2.662	-1.923
Bahrain	-1.905	-2.703	-1.906	-1.097
Cambodia	-4.939	-5.883	-4.936	-4.002
Israel	-1.833	-2.630	-1.835	-1.034
Malaysia	-2.759	-3.529	-2.757	-2.002
Mongolia	-1.755	-2.632	-1.753	-0.898
Saudi Arabia	-4.173	-5.028	-4.177	-3.285
Sri Lanka	-3.280	-4.077	-3.283	-2.474
Thailand	-3.195	-3.970	-3.197	-2.422
Turkey	-2.328	-3.059	-2.328	-1.601
Belgium	-1.529	-2.275	-1.530	-0.779
Bulgaria	-2.407	-3.172	-2.404	-1.662
Czechia	-1.927	-2.640	-1.927	-1.206
Denmark	-1.995	-2.761	-1.996	-1.237
Estonia	-1.958	-2.663	-1.958	-1.255
France	-1.584	-2.340	-1.583	-0.826
Greece	-1.883	-2.682	-1.886	-1.065
Ireland	-1.537	-2.289	-1.538	-0.783
Italy	-2.275	-3.020	-2.274	-1.546
Latvia	-1.926	-2.708	-1.925	-1.147
Netherlands	-1.972	-2.706	-1.970	-1.257
Norway	-2.058	-2.807	-2.055	-1.326
Poland	-2.080	-2.809	-2.076	-1.358
Russia	-2.334	-3.133	-2.336	-1.525
Slovenia	-2.084	-2.824	-2.082	-1.369
Switzerland	-1.862	-2.592	-1.860	-1.143
United Kingdom	-1.901	-2.656	-1.904	-1.138
Canada	-2.769	-3.649	-2.772	-1.876
Dominican Republic	-3.376	-4.227	-3.375	-2.525
Jamaica	-3.416	-4.284	-3.415	-2.559
Argentina	-2.099	-2.989	-2.098	-1.216
Chile	-2.286	-3.201	-2.290	-1.344
Ecuador	-3.509	-4.407	-3.504	-2.640
Uruguay	-2.214	-3.069	-2.213	-1.369

Table 4: Non-SL Model Predicting Result

	mean	2.5%	50%	97.5%
Kenya	-4.891	-5.744	-4.898	-3.997
Tunisia	-2.836	-3.796	-2.829	-1.908
Zimbabwe	-4.463	-5.284	-4.463	-3.633
Azerbaijan	-2.619	-3.384	-2.617	-1.868
Bahrain	-1.876	-2.690	-1.876	-1.053
Cambodia	-4.981	-5.928	-4.982	-4.041
Israel	-1.782	-2.579	-1.784	-0.960
Malaysia	-2.770	-3.547	-2.769	-2.005
Mongolia	-1.708	-2.588	-1.705	-0.841
Saudi Arabia	-4.062	-4.947	-4.069	-3.137
Thailand	-3.156	-3.939	-3.155	-2.382
Turkey	-2.303	-3.038	-2.302	-1.563
Belgium	-1.510	-2.262	-1.511	-0.753
Bulgaria	-2.397	-3.178	-2.394	-1.643
Czechia	-1.916	-2.638	-1.918	-1.187
Denmark	-1.993	-2.755	-1.992	-1.230
Estonia	-1.964	-2.674	-1.963	-1.258
France	-1.563	-2.311	-1.563	-0.808
Greece	-1.888	-2.701	-1.893	-1.067
Ireland	-1.540	-2.285	-1.541	-0.787
Italy	-2.277	-3.012	-2.276	-1.546
Latvia	-1.940	-2.728	-1.941	-1.151
Netherlands	-1.986	-2.719	-1.985	-1.252
Norway	-2.067	-2.820	-2.066	-1.323
Poland	-2.067	-2.808	-2.065	-1.332
Russia	-2.333	-3.147	-2.333	-1.516
Slovenia	-2.089	-2.826	-2.085	-1.363
Switzerland	-1.870	-2.598	-1.869	-1.140
United Kingdom	-1.894	-2.655	-1.895	-1.124
Canada	-2.551	-3.367	-2.550	-1.725
Dominican Republic	-3.338	-4.185	-3.339	-2.479
Jamaica	-3.368	-4.245	-3.365	-2.506
United States	-2.087	-2.928	-2.085	-1.264
Argentina	-2.086	-2.993	-2.082	-1.193
Chile	-2.295	-3.221	-2.301	-1.336
Ecuador	-3.528	-4.424	-3.524	-2.658
Uruguay	-2.199	-3.051	-2.201	-1.342