Regression Models to Analyze COVID-19 Impact on 163 Countries

Class Project – Regression Analysis (563)

DECEMBER 12, 2021

RUTGERS UNIVERSITY
Guided by – Prof. Jack Mardekian

Table of Contents

ABSTRACT

The data obtained from Ninth World Happiness Report (2021). Which focuses on the effects of COVID-19 and how people all over the world have fared. There are two main objectives of this project. The data contains observations from over 160 countries for 11 demographic and economic variables. First, to study relation between deaths due to COVID-19 and certain demographic factors pertaining to the countries. Second, to try to contrast and compare between the countries with fewer deaths per 100,000 population to the countries with higher number of deaths per 100,000 population. I have used multiple linear regression model to achieve first objective. Stepwise regression followed by K-fold cross validation was used for variable selection. I have used logistic regression to classify countries in "high" and "low" brackets based on the number of deaths in the country. There is a significant relationship between the most variables and the deaths per hundred thousand population. The result suggest that the variables under consideration were able to predict around 60% of the variation in the deaths variable.

INTRODCUTION

The dataset measures the effect of COVID-19 on 167 countries using 17 different variables. We are using deaths per hundred thousand people as a dependent variable to perform multiple regression hoping to quantify effect of other variables on the adverse effects of the pandemic. For categorization, observations are assigned groups, “low” and “high, based on the number of deaths in the country per hundred thousand of population.

We are analyzing demographic features like population and median age to understand how they relate to increase/decrease in the deaths due to COVID-19. For analyzing the economic drivers behind the effective handling of the pandemic we are considering features such as Gini coefficient of income for the country, if the country is being headed by a woman, exposure to the COVID-19 in the other countries, etc.

MATERIALS AND METHODS

1. Dataset:

This dataset contains data such as the various countries selected, their respective populations in the year 2019 and 2020, total COVID-19 deaths in each of these countries, the median age in the corresponding countries considered, whether it is an island or not, index of exposure to COVID-19 infections in other countries as of march 31, log of average distance to SARS countries, whether it is a WHO western pacific region, whether a female head of government or not. In particular, we try to explain why some countries have done so much better than others. The detailed dataset can be found in the following website here with some other interesting statistics!

Following are the descriptions for each of the features we considered here for the analysis:

Population.2020 : numeric variable containing 2020 population of the country

Median.age: numeric variable containing median age of the country as of 2020

Island: factor variable representing of the country is an island or not

Index.of.insitutional.trust: numeric variable representing the people's trust in their respective public institutions

Gini.coefficient.of.income: numeric variable representing Gini coefficient (a measure of the distribution of income across a population). More about it can be found here.

Country: character variable

Deaths.per.100k: numeric variable representing people died per 100,000 population

log.avg.dist.SARS: numeric variable representing log of average distance from SARS countries

WPAC: factor variable representing if the country is a WHO Western Pacific region on not

Female.lead: factor variable representing if the country has a female head or not

Exposure.index.other: numeric variable representing the index of exposure of COVID-19 infections in the corresponding country

death.brackets: factor variable, created just for this analysis (not a part of the original data), representing if the country has "low" or "high" levels of deaths

2. Methods:

The raw data was processed to arrive at a cleaner dataset as follows: after renaming the variables for better readability, last five variables are removed from the analysis because they contain more than 100 missing values. “Population 2019” is removed as it is highly correlated with “Population 2020” and using either of them would have similar effect. The analysis would have been more useful if we kept “Excess deaths in 2020” to better quantify the effects, however, even that variable has most of its values missing.

The variables under consideration have various different units that affects their absolute values greatly. Thus, variables are centered and scaled for better homogeneity and interpretability of the results.

Three countries, Turkmenistan, Somalialand and North Cyprus have their population missing. For Somaliland and North Cyprus, most of the other variables values are also missing. So, removing these two countries from the analysis. For Turkmenistan, median age as well population are missing. After a bit of research it was found that WHO does not have COVID 19 cases/deaths data for Turkmenistan. Rather Turkmenistan claims to have no cases/deaths. Thus, removing Turkmenistan from the analysis.

In the end, we have 163 observations of 12 variables of interest.

Next, we fit a multiple regression model followed by using it to perform forward, backward, stepwise bidirectional, all possible, and finally best subset regression using the “olsrr” package. We have used p values to arrive at the best model in each of the above multiple regressions. Then, we perform repeated K-fold cross-validation for better performance on the test data set.

A brief summary of different theoretical concepts used-

Forward selection: This method only adds one variable per step (starting with a null model) to the model if the model improves a chosen model performance criteria.

Backward elimination: This method only removes one variable per step (starting from a full model) from the model to improve a chosen model performance criteria.

Stepwise bi-directional: In this method, variables are added as well as removed from the model as long as it improves the chosen model performance criteria.

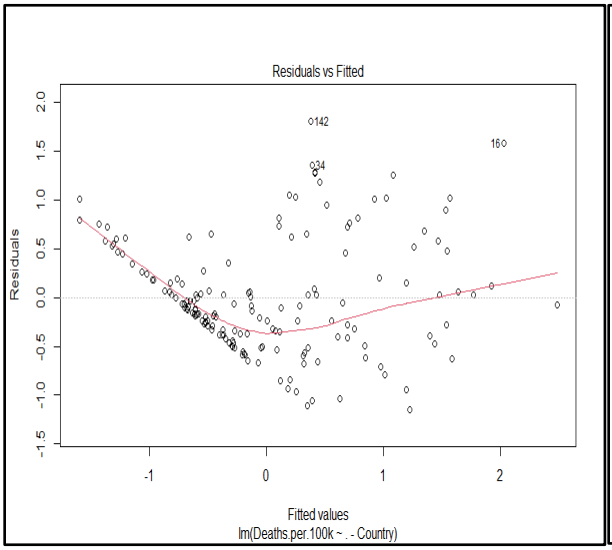
Best subset: This method chooses best subset of variables based on the chosen model performance criteria for each number of variables (one to total number of variables).

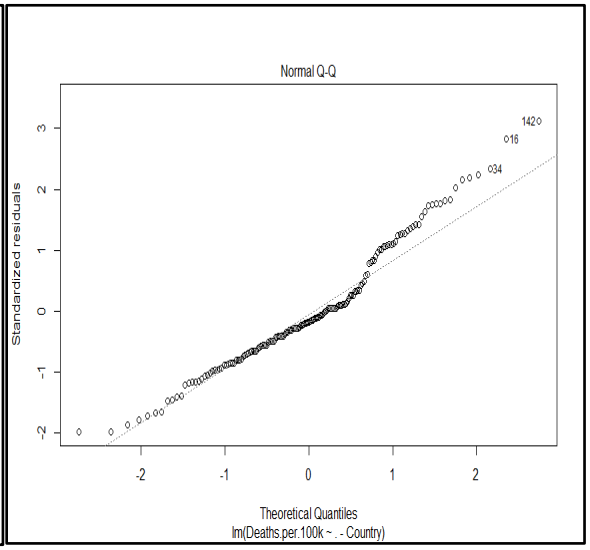
Repeated K-fold cross-validation: In K-fold cross-validation, the dataset is randomly split into K-folds of equal size. One of the folds is used as a validation set and others as a training set. This process is repeated to fine tune the results. We have chosen K=10 and repeated cross-validation 6 times for each of the model selection methods.

RESULTS

1. OLS Regression: This is a full OLS model with deaths per hundred thousand of population as a dependent variable. All the variables except population 2020 and WPAC are highly significant.

```
Call:
lm(formula = Deaths.per.100k ~ . - Country, data = WHR)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1492 -0.3839 -0.1030  0.3076  1.8081

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    0.14196    0.05684   2.498 0.013557 *
Population.2020                0.01550    0.04923   0.315 0.753335
Median.age                     0.44754    0.07074   6.327 2.63e-09 ***
Island1                       -0.40447    0.14591  -2.772 0.006262 **
Index.of.institutional.trust  -0.21723    0.05115  -4.247 3.75e-05 ***
Gini.coefficient.of.income     0.16523    0.06183   2.672 0.008349 **
log.avg.dist.SARS              0.16268    0.06555   2.482 0.014157 *
WPAC1                         -0.22848    0.21835  -1.046 0.297041
Female.lead1                  -0.48012    0.13880  -3.459 0.000702 ***
Exposure.index.other           0.44204    0.07438   5.943 1.82e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5908 on 153 degrees of freedom
Multiple R-squared:  0.6703,    Adjusted R-squared:  0.6509
F-statistic: 34.57 on 9 and 153 DF,  p-value: < 2.2e-16
```

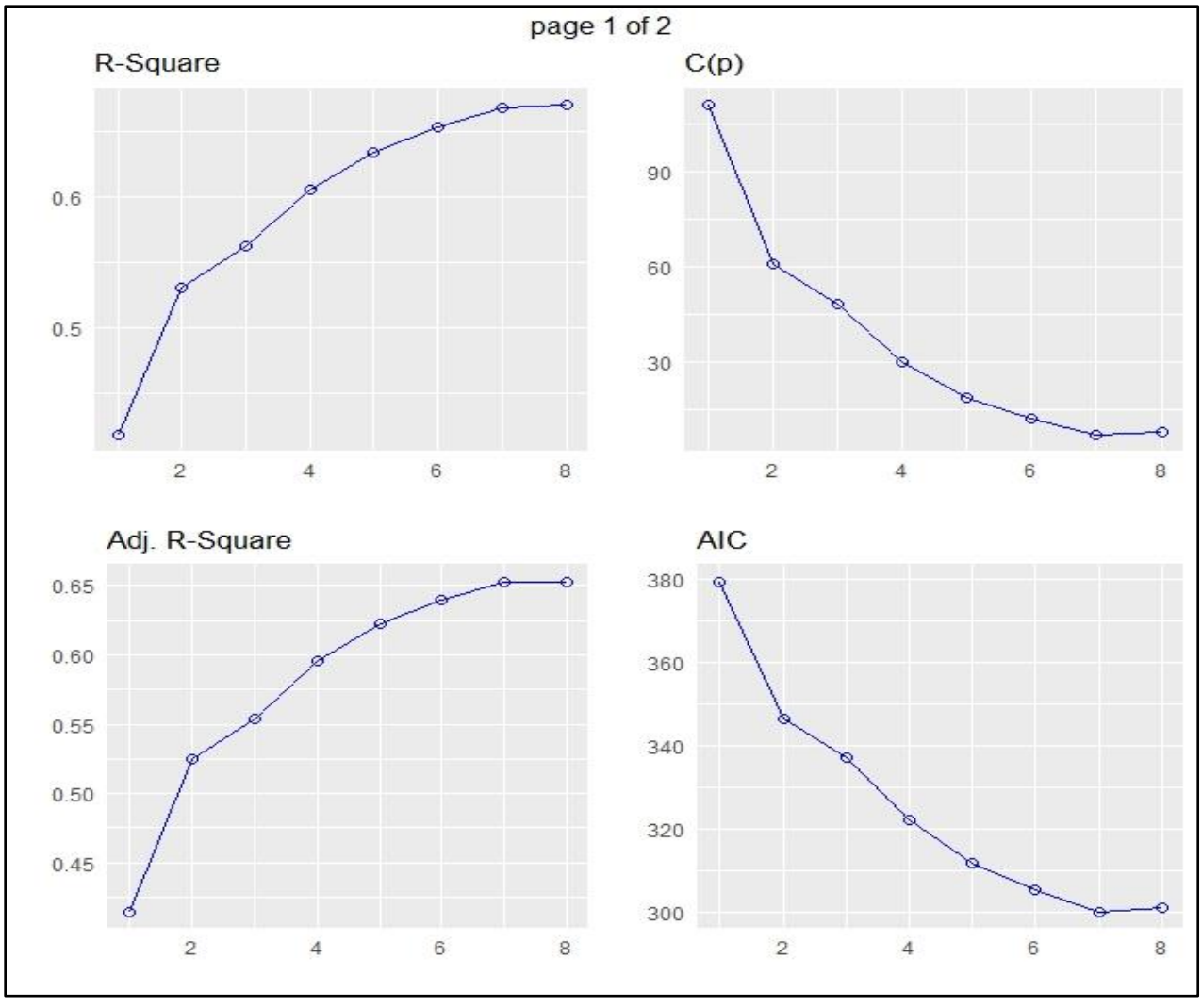




2. Forward Selection: The summary shows increase in Cp and AIC after adding the last variable WPAC to the model.

Selection Summary

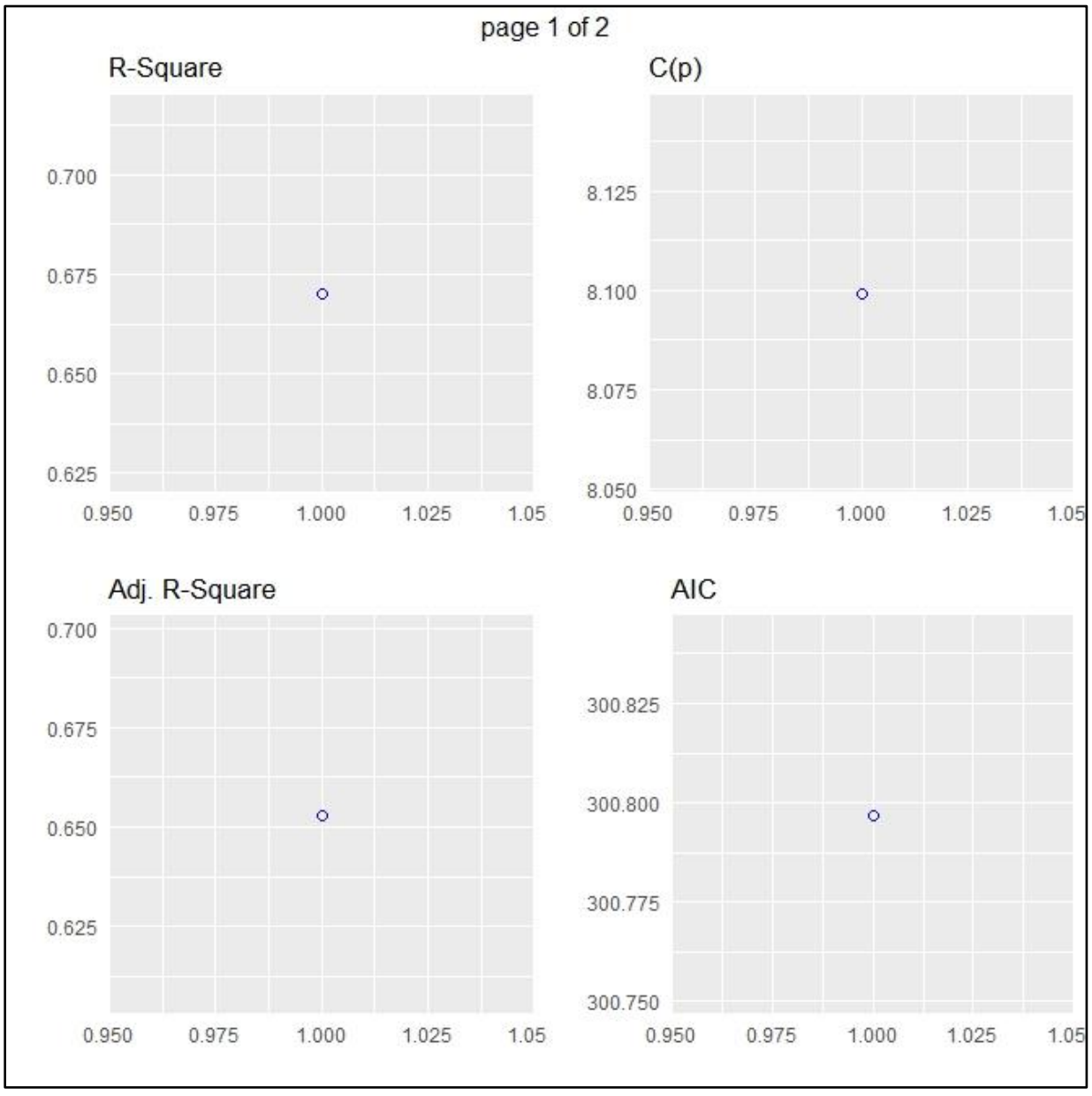
Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Exposure.index.other	0.4180	0.4144	111.0986	379.3323	0.7652
2	Index.of.institutional.trust	0.5307	0.5249	60.7888	346.2447	0.6893
3	Median.age	0.5622	0.5540	48.1724	336.9209	0.6679
4	log.avg.dist.SARS	0.6055	0.5955	30.0800	321.9474	0.6360
5	Female.lead	0.6338	0.6221	18.9675	311.8339	0.6147
6	Island	0.6525	0.6392	12.2714	305.2734	0.6007
7	Gini.coefficient.of.income	0.6678	0.6528	7.1910	299.9552	0.5893
8	WPAC	0.6701	0.6530	8.0991	300.7969	0.5891



3. Backward Elimination: Only the variable "Population 2020" is eliminated.

Elimination Summary

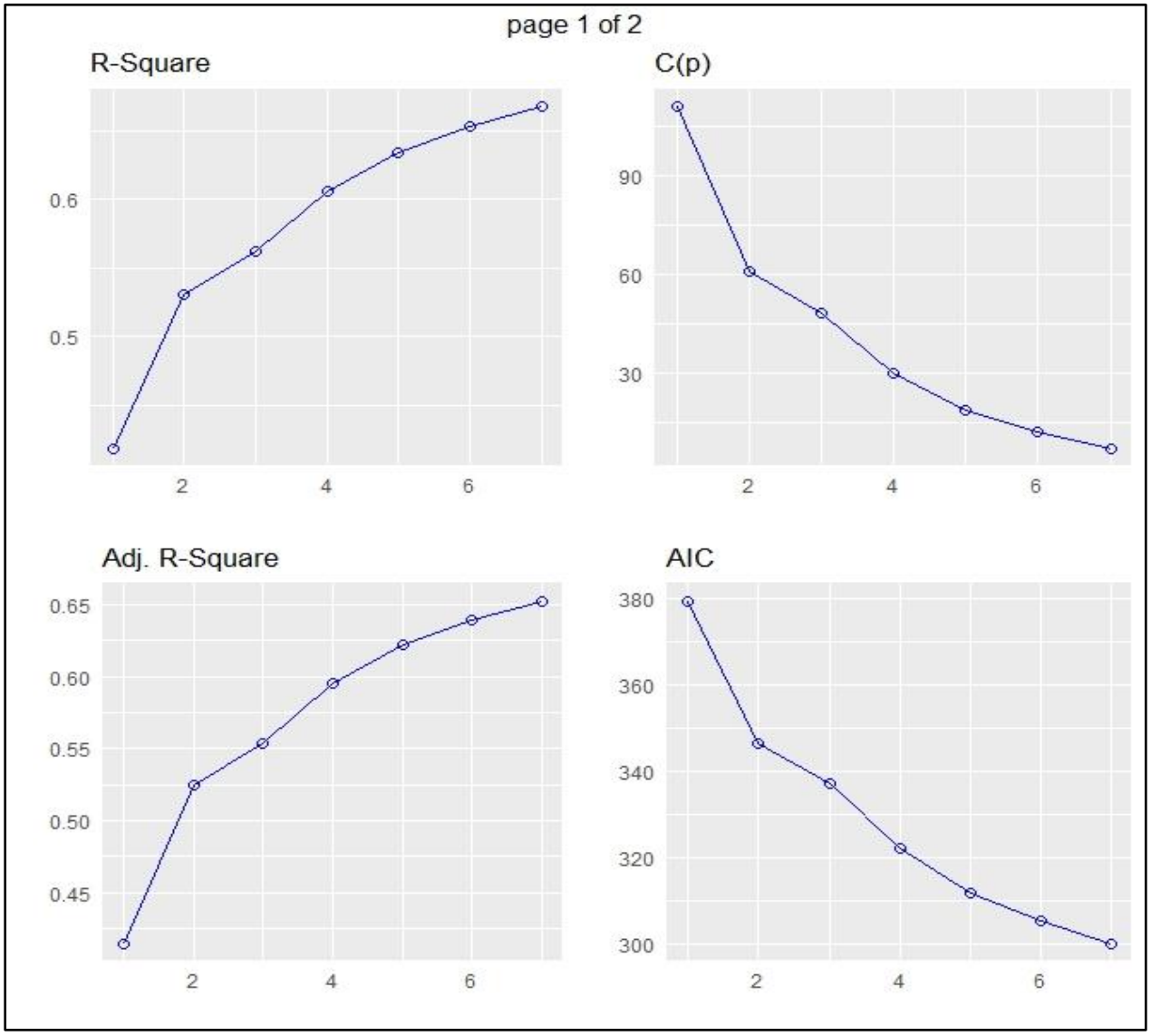
Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Population.2020	0.6701	0.653	8.0991	300.7969	0.5891



4. Stepwise Bidirectional: The stepwise both ways included 7 predictors in its final model.

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Exposure.index.other	addition	0.418	0.414	111.0990	379.3323	0.7652
2	Index.of.institutional.trust	addition	0.531	0.525	60.7890	346.2447	0.6893
3	Median.age	addition	0.562	0.554	48.1720	336.9209	0.6679
4	log.avg.dist.SARS	addition	0.606	0.596	30.0800	321.9474	0.6360
5	Female.lead	addition	0.634	0.622	18.9680	311.8339	0.6147
6	Island	addition	0.653	0.639	12.2710	305.2734	0.6007
7	Gini.coefficient.of.income	addition	0.668	0.653	7.1910	299.9552	0.5893



5. Best Subset: From the list, the best best subset is the one with 7 predictor variables.

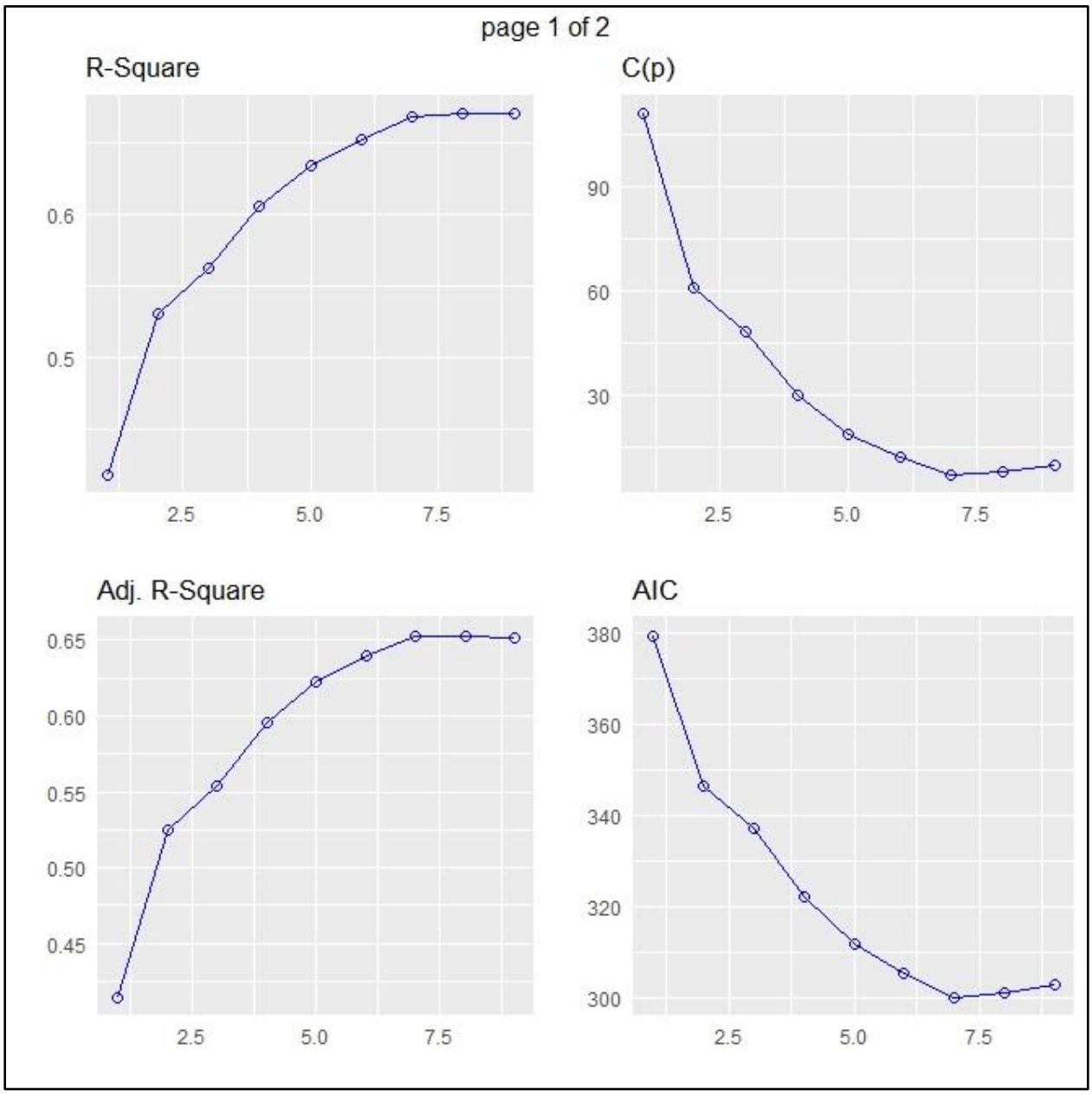
Best Subsets Regression

Model Index	Predictors
1	Exposure.index.other
2	Index.of.institutional.trust Exposure.index.other
3	Median.age Index.of.institutional.trust Exposure.index.other
4	Median.age Index.of.institutional.trust log.avg.dist.SARS Exposure.index.other
5	Median.age Index.of.institutional.trust log.avg.dist.SARS Female.lead Exposure.index.other
6	Median.age Island Index.of.institutional.trust log.avg.dist.SARS Female.lead Exposure.index.other
7	Median.age Island Index.of.institutional.trust Gini.coefficient.of.income log.avg.dist.SARS Female.lead Exposure.index.other
8	Median.age Island Index.of.institutional.trust Gini.coefficient.of.income log.avg.dist.SARS WPAC Female.lead Exposure.index.other
9	Population.2020 Median.age Island Index.of.institutional.trust Gini.coefficient.of.income log.avg.dist.SARS WPAC Female.lead Exposure.index.other

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.4180	0.4144	0.4008	111.0986	379.3323	NA	388.6135	95.4499	0.5928	0.0037	0.5964
2	0.5307	0.5249	0.5109	60.7888	346.2447	NA	358.6197	77.4483	0.4839	0.0030	0.4869
3	0.5622	0.5540	0.5372	48.1724	336.9209	NA	352.3897	72.7078	0.4570	0.0028	0.4598
4	0.6055	0.5955	0.5768	30.0800	321.9474	NA	340.5099	65.9348	0.4169	0.0026	0.4194
5	0.6338	0.6221	0.6021	18.9675	311.8339	NA	333.4902	61.6048	0.3918	0.0024	0.3942
6	0.6525	0.6392	0.6175	12.2714	305.2734	NA	330.0234	58.8300	0.3763	0.0023	0.3787
7	0.6678	0.6528	0.6307	7.1910	299.9552	NA	327.7990	56.6124	0.3643	0.0023	0.3665
8	0.6701	0.6530	0.6282	8.0991	300.7969	NA	331.7344	56.5789	0.3662	0.0023	0.3684
9	0.6703	0.6509	0.6211	10.0000	302.6913	NA	336.7226	56.9143	0.3705	0.0023	0.3728

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp
APC: Amemiya Prediction Criteria


page 1 of 2
R-Square
0.6
0.5
2.5
5.0
7.5
C(p)
90
60
30
2.5
5.0
7.5
Adj. R-Square
0.65
0.60
0.55
0.50
0.45
2.5
5.0
7.5
AIC
380
360
340
320
300
2.5
5.0
7.5

6. Repeated 10-fold cross-validation:

a) Forward Selection:

```
> print(cv_fwd)
Linear Regression

163 samples
  8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 146, 147, 147, 147, 147, 147, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5959861  0.6676237  0.471286

Tuning parameter 'intercept' was held constant at a value of TRUE
```

b) Backward Elimination:

```
> print(cv_back)
Linear Regression

163 samples
 11 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 147, 147, 147, 147, 147, 147, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.788133  0.1846253  1.315842

Tuning parameter 'intercept' was held constant at a value of TRUE
```

c) Stepwise Bi-directional:

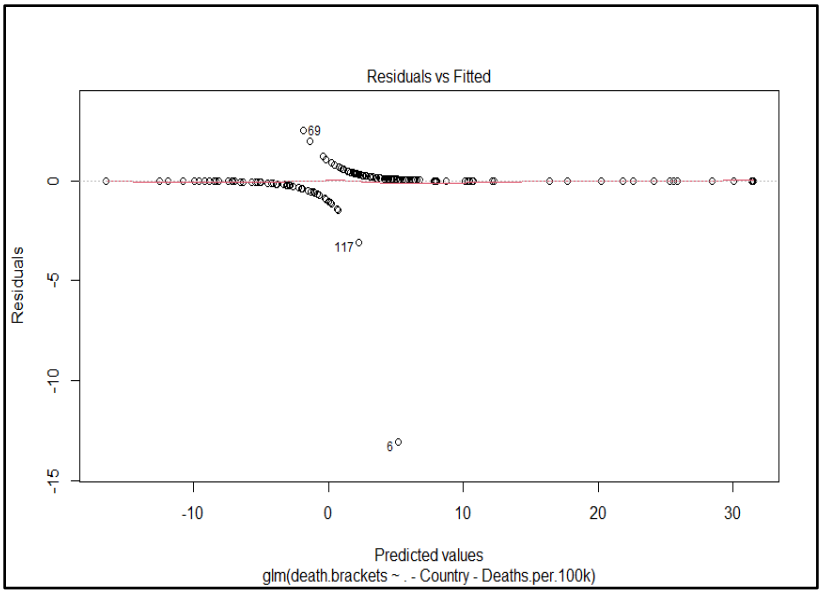
```
> print(cv_both)
Linear Regression

163 samples
  7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 147, 147, 147, 146, 147, 147, ...
Resampling results:

  RMSE      Rsquared   MAE
  0.596177  0.6547217  0.4706217

Tuning parameter 'intercept' was held constant at a value of TRUE
```

d) Best Subset:

```
> print(cv_best)
Linear Regression

163 samples
  7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 147, 147, 147, 147, 147, 147, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5955802  0.6519634  0.4726814

Tuning parameter 'intercept' was held constant at a value of TRUE
```

7. Logistic Regression:

```
> summary(WHR_logit)

Call:
glm(formula = death.brackets ~ . - Country - Deaths.per.100k,
    family = "binomial", data = WHR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2086  -0.0445   0.0257   0.1736   2.0059

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                     0.6409     0.4782   1.340 0.180093
Population.2020                -0.2230     1.0373  -0.215 0.829754
Median.age                     -2.3572     0.6480  -3.638 0.000275 ***
Island1                         2.6306     1.0222   2.573 0.010071 *
Index.of.institutional.trust    2.3298     0.6654   3.501 0.000463 ***
Gini.coefficient.of.income     -1.7074     0.5696  -2.998 0.002722 **
log.avg.dist.SARS              -0.6043     0.5444  -1.110 0.267018
WPAC1                          19.7071  1951.2307   0.010 0.991942
Female.lead1                    3.3162     1.2750   2.601 0.009295 **
Exposure.index.other           -3.5658     1.1131  -3.204 0.001358 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 211.006  on 162  degrees of freedom
Residual deviance:  60.466  on 153  degrees of freedom
AIC: 80.466

Number of Fisher Scoring iterations: 18
```


Residuals vs Fitted
69
117
6
Residuals
0
-5
-10
-15
-10
0
10
20
30
Predicted values
glm(death.brackets ~ . - Country - Deaths.per.100k)

DISCUSSION

The OLS full model has an adjusted R-sqrd of 0.6509 and it includes all the variables except "Population 2020" and "WPAC" as significant. Forward selection and Backward elimination models confirm the removal of Population 2020 from the model as it is not included in either of them. Both these methods basically gave the same model with an adjusted R-sqrd of 0.653, slightly better than the OLS full model. Stepwise bidirectional selects all the features that forward/backward models selected except for the feature "WPAC". However, the adjusted R-sqrd is almost the same (0.653). The best subset regression gives best subsets of features for different number of total features used in the model. The 7 feature best subset is same as the one given bi bidirectional stepwise regression.

The logistic regression includes "median.age", "Island1", "Index.of.insitutional.trust", "Gini.coefficient.of.income", "Female.lead1", and "Exposure.index.other" as significant variables.

Interpretation of the above results:

Strictly going by the selected model and its beta parameters, a country with low median age, which is an island, which has high institutional trust, is far from the SARS countries, has high Gini index, which is headed by a female should fare better when it comes to COVID-19 deaths irrespective of its population. This interpretation, however, comes with the following caveats.

Based on the multiple regression analysis, one of the more surprising findings is that the effect of the gender of the country head on the number of deaths per hundred thousand people is almost same as that of median age. However, the data is not even on this feature, that is, only 23 of the 163 countries under consideration have governments headed by female. So, this significance doesn't mean much in terms of interpretation. Same is true for a country being an island/not an island.

Another interesting but rather obvious finding is the negative linear relationship between institutional trust and deaths. It is only to be expected that more trusted governments take better measures to ensure the lesser deaths, although that is not always necessary.

Coming to the classification, significant predictors are same as what we got in the best subset model except the distance of the country from SARS countries. Looking at the change in odds table, we can see that among the most significant variables, institutional index has the most effect towards changing odds per unit change.

This analysis helps us understand the trends in COVID-19 trends better by tying it with the significant variables. We understood that population is not at all a significant factor while age and institutional trust are very important factors along with Gini index.

ACKNWOLEDGMENTS

The data for the analysis is obtained from Kaggle. Kaggle in turn obtained this data from the World Happiness Report. It is part of the World Happiness Report 2021. The original dataset, its full description, and other statistics can be found here in detail.