COVID-19 Data Analysis

Trevor Carpenter, Christina De Cesaris, Michelle Tran
STA 135: Multivariate Data Analysis
Professor Maxime Pouokam
University of California, Davis
June 7, 2021

I. Introduction

The COVID-19 pandemic is an event that will have long-lasting effects in our society. While no country was entirely prepared for the chaos that ensued, there were definitely some countries that had a better handle on the situation than others. Countries who were quick on putting out preventative measures and were able to produce masks quickly were able to keep their population's total cases to a minimum, whereas countries who lagged behind on safety measures quickly saw the exponent rise in cases and the overcapacity of hospitals.

As we are nearing the end of pandemic, much data has been collected and analyzed. In our paper we have decided to look at data involving the mortality recovery ratio in regards to the beginning of the pandemic. [insert question of interest and analysis technique]

II. Data, Models, and Methods

The COVID-19 data used here is publicly and available from Worldometer website https://www.worldometers.info/coronavirus/ for March 30, April 15, and April 25, 2020. Data were captured on the next day to these specified dates. Countries with COVID-19 total cases less than 500 or countries with missing data were omitted from the analysis to keep good representability of each variable. Number of countries included in the analysis was 56 countries on March 30, 82 countries on April 15, and 91 countries on April 25.

The variables included; in any given country, total cases refers to total cases confirmed with COVID-19; active cases refers to total number of open cases (mild, serious, or critical); total deaths refers to total deaths with COVID-19; critically ill cases refers to number of serious/critically ill cases; mortality recovery ratio refers to the ratio between total deaths to total recovered patients.

III. Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3038600	0.1280644	2.3727127	0.0390958
TotalCases	0.0000426	0.0000830	0.5136285	0.6186686
TotalDeaths	0.0051524	0.0001694	30.4150162	0.0000000
ActiveCases	-0.0008670	0.0001057	-8.2021353	0.0000095
Critical	0.0244141	0.0008389	29.1038920	0.0000000

```
## Start:  AIC=-29.97
## MortalityRecoveryRatio ~ TotalCases + TotalDeaths + ActiveCases +
##     Critical
##
##               Df Sum of Sq    RSS     AIC
## - TotalCases   1     0.028  1.072 -31.579
## <none>                      1.044 -29.969
## - ActiveCases  1     7.026  8.070  -1.298
## - Critical     1    88.463 89.507  34.794
## - TotalDeaths  1    96.613 97.657  36.101
##
## Step:  AIC=-31.58
## MortalityRecoveryRatio ~ TotalDeaths + ActiveCases + Critical
##
##               Df Sum of Sq    RSS     AIC
## <none>                        1.1 -31.579
## - TotalDeaths  1    1689.3 1690.3  76.869
## - ActiveCases  1    3760.4 3761.5  88.868
## - Critical     1    4464.6 4465.6  91.442
```

```
##
## Call:
## lm(formula = MortalityRecoveryRatio ~ TotalDeaths + ActiveCases +
##     Critical, data = covidpt)
##
## Coefficients:
## (Intercept)  TotalDeaths  ActiveCases     Critical
##   0.3487820    0.0052368   -0.0008128    0.0239874
```

Based on both the step model and the linear regression model, we see that it is likely that the Total Cases variable is not a good predictor of the Mortality Recovery Ratio. However this could be misleading due to correlation between variables, so we can check with a full and reduced ANOVA model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3487820	0.0903634	3.85977	0.0026544
TotalDeaths	0.0052368	0.0000398	131.66213	0.0000000
ActiveCases	-0.0008128	0.0000041	-196.44007	0.0000000
Critical	0.0239874	0.0001121	214.04372	0.0000000

We build the model based on the following variables:

$Y = MortalityRecoveryRatio$
$X_1 = TotalCases$
$X_2 = TotalDeaths$
$X_3 = ActiveCases$
$X_4 = Critical$

Our reduced model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ and our full model is $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

Running the anova model we get the following.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
10	1.044377	NA	NA	NA	NA
11	1.071930	-1	-0.0275522	0.2638143	0.6186686

Interestingly the P-value for the ANOVA model is quite high and the F-value is very low, so we cannot conclude that there is any statistically significant difference between the reduced and full models in predicting the Y value.

IV. Conclusion & Future Work

Appendix: R Script

```r
knitr::opts_chunk$set(echo = F, warning = F, message = F)
library(knitr)
library(dplyr)
library(readr)
covid = as.data.frame(read_delim("data/COVID.csv", delim = ','))
covidpt = covid[-c(4, 10, 16), 3:7]

fmodel = lm(MortalityRecoveryRatio ~., data = covidpt)
summary(fmodel)$coefficients %>% kable()
step(fmodel)
rmodel = lm(MortalityRecoveryRatio ~ TotalDeaths + ActiveCases
            + Critical, data = covidpt)
summary(rmodel)$coefficients %>% kable()
anova(fmodel,rmodel) %>% kable()
```