STA 135 Final Project

Angela Guo, Xinmei Wang, Chunqiu Li

6/7/2021

I. Introduction

Coronavirus is a large group of viruses that may cause human diseases. Since the outbreak of the new coronavirus pandemic in late 2019, governments around the world have taken different measures to deal with this epidemic. Although it has been more than a year since the outbreak began, the pandemic is still a serious problem. Therefore, it is important to apply analysis to find possible solutions to prevent the pandemic from getting even more serious. In this report, we will apply Multivariate Analysis of Variance to our dataset to compare the means from three differents dates.

II. Data, Models, and Methods

```
##    Time Parameter Total.Cases Total.Deaths Active.Cases Critical
## 1 30Mar       Min       515.0         1.00         99.0      1.0
## 2 30Mar    1st Qu       954.8         9.75        866.8     11.0
## 3 30Mar    Median      1937.0        26.50       1867.5     38.0
## 4 30Mar      Mean     12789.1       611.55       9416.5    499.1
## 5 30Mar    3rd Qu      7014.8       122.00       4628.2    170.5
## 6 30Mar       Max    142746.0     10779.00     135695.0   5231.0
##   Mortality.Recovery.Ratio
## 1                   0.0100
## 2                   0.0775
## 3                   0.3350
## 4                   3.1705
## 5                   1.3250
## 6                  72.0000
```

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.

We are going to use Multivariate Analysis of Variance (MANOVA) for that dataset. The purpose here is to determine if there are differences in the means of different statistics for those three dates. Let $Y_{ij}, j = 1, \ldots, m_i$, be i.i.d. $N_p(u_i, V), i = 1, \ldots, 3$. We write the one factor Manova model as $Y_{ij} = \mu_i + \epsilon_{ij}$, where $\mu = \Sigma(m_i/n)\mu_i$ and $\alpha_i = \mu_i - \mu$.

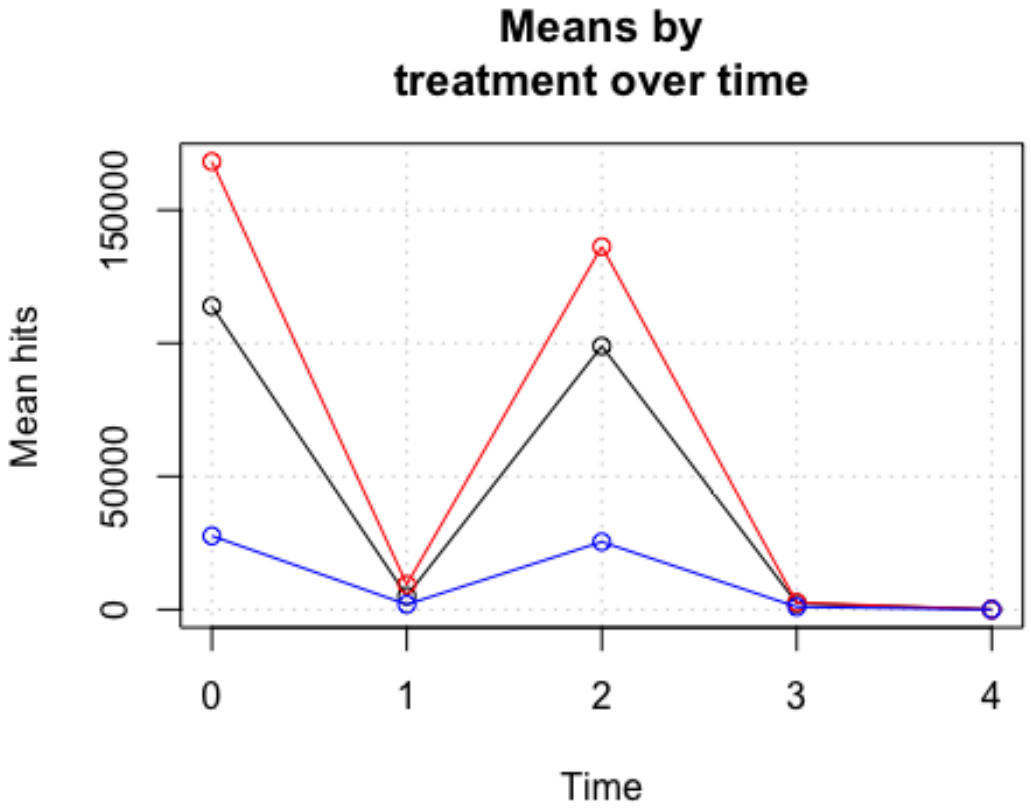
Our goal is to test the nullhypothesis against the alternative hypothesis:

$H_0: \mu_{mar30} = \mu_{apr15} = \mu_{apr25}$

H_a: Not all equal

Before that let's look at the vector mean we are comparing.

```
##    Time Total.Cases Total.Deaths Active.Cases  Critical
## 1 15Apr   113923.33     5078.112     98909.67 2387.4600
## 2 25Apr   168204.33     9462.650    136270.33 2652.7833
## 3 30Mar    27659.45     1924.967     25428.83  991.7667
##   Mortality.Recovery.Ratio
## 1                 2.292633
## 2                 1.155367
## 3                12.819667
```



Now we perform model fitting:

```
## Call:
##    manova(formula = cbind(Total.Cases, Total.Deaths, Active.Cases,
##     Critical, Mortality.Recovery.Ratio) ~ Time, data = data)
##
## Terms:
##                                    Time    Residuals
## Total.Cases                6.028150e+10 1.107809e+12
## Total.Deaths                  171966340   3166217480
## Active.Cases                38162180807 787895657768
## Critical                        9554665    356316537
## Mortality.Recovery.Ratio            496         4368
## Deg. of Freedom                       2           15
##
## Residual standard errors: 271760.8 14528.63 229186.3 4873.852 17.06396
## Estimated effects may be unbalanced
```

III. Insignificant Differences

By computing Wilks test, the P-value that we got is 0.7544 which is very large and not significant. Therefore, based on the Wilks test, we do not reject the null hypothesis and conclude that the means for different stats on those three dates are the same.

```
##           Df   Wilks approx F num Df den Df Pr(>F)
## Time       2 0.59486  0.65243     10     22 0.7544
## Residuals 15
```

Since the P-value from the Wilks test is very large, we want to apply the Pillai test to see if the conclusion that we get will be similar. After computing, the p-value that we have received is 0.7096 which is also very large. Therefore, we reject the null hypothesis again.

```
##           Df  Pillai approx F num Df den Df Pr(>F)
## Time       2 0.45492  0.70663     10     24 0.7096
## Residuals 15
```

After computing the Roy test, the p-value that we have received is 0.5143 which is smaller compared to the previous two tests. However, the value is still insignificant. Therefore, we reject the null hypothesis again.

```
##           Df     Roy approx F num Df den Df Pr(>F)
## Time       2 0.37317  0.89562      5     12 0.5143
## Residuals 15
```

IV. Conclusion

We performed three ways of Manova (Wilks, Pillai and Roy) on testing the null hypothesis of the means of the covid stats on the three dates (Mar 30th, Apr 15th and Apr 30th) are the same. The results that we have gotten from the computation show that there is not enough evidence to reject the null hypothesis. Therefore, we conclude that the means of the covid stats from those three dates are the same.

Appendix Code

```
data <- read.csv("/Users/angelaguo/Desktop/COVID.data.csv")
head(data)
mar_30 <- subset(data, Time == "30Mar")
apr_15 <- subset(data, Time == "15Apr")
apr_25 <- subset(data, Time == "25Apr")
mu_totalcases <- c(mean(mar_30$Total.Cases), mean(apr_15$Total.Cases),
mean(apr_25$Total.Cases))
mu_totaldeaths <- c(mean(mar_30$Total.Deaths), mean(apr_15$Total.Deaths),
mean(apr_25$Total.Deaths))
mu_activecases <- c(mean(mar_30$Active.Cases), mean(apr_15$Active.Cases),
mean(apr_25$Active.Cases))
mu_critical <- c(mean(mar_30$Critical), mean(apr_15$Critical),
mean(apr_25$Critical))
mu_mortality <- c(mean(mar_30$Mortality.Recovery.Ratio),
mean(apr_15$Mortality.Recovery.Ratio), mean(apr_25$Mortality.Recovery.Ratio))
save.means<-aggregate(formula = cbind(Total.Cases, Total.Deaths,
Active.Cases, Critical, Mortality.Recovery.Ratio) ~ Time, data = data, FUN =
mean)
save.means
plot(x = 0:4, save.means[1,-1], main = "Means by
treatment over time", ylim = c(min(save.means[,-1]),
max(save.means[,-1])), panel.first = grid(), type =
"o", col = "black", xlab = "Time", ylab = "Mean hits")
lines(x = 0:4, save.means[2,-1], type = "o", col = "red")
lines(x = 0:4, save.means[3,-1], type = "o", col =
"blue")
legend(x = 0, y = 94, legend =
levels(as.factor(data$Time)), col = c("black", "red",
"blue"), lty = 1, bty =
"n")
model<-manova(formula = cbind(Total.Cases, Total.Deaths, Active.Cases,
Critical, Mortality.Recovery.Ratio) ~ Time, data = data)
model
# different MANOVA TESTs
summary(model, test = "Wilks")
summary(model, test = "Pillai")
summary(model, test = "Roy")
```