COVID-19 Case Study - Midterm

your name

Instruction

All the teaching team members will be available from 7:00 - 9:10 PM. The submission will be closed sharp at 9:10PM.

Instruction: This midterm requires you to use R. It is completely open book/notes/internet. Write your answers using .rmd format and knitr it into one of the html/pdf/docx format. Show your codes, plots or R-output when needed. You can use `echo = TRUE` to show your codes which is the default setup for this file. If you have trouble formatting the plots, don't worry about it. We are not looking for pretty solutions, rather to see if you are able to make sense out of the data using R. Make sure the compiled pdf/html/docx (only one of them) shows your answers completely and that they are not cut-off. Throughout the exam, you do not need to use any LaTeX or mathematical equations. **Whenever we ask for test at some significant level, assume all the model assumptions are satisfied.**

All the answers should be clearly supported by relevant R code or based on the R output.

There are 4 questions with various parts:

- **Question 1: 3 parts**
- **Question 2: 4 part**
- **Question 3: 1 parts**
- **Question 4: 9 parts**

Data needed for the Midterm: `/canvas/Files/Exams/Midterm/Midterm Spring 2022/data/covid_county_midterm_s`

Electronic Submission: Two files needed: your `.rmd` file and a compiled file (either a pdf/html/docx). **Label them with your full name.** In the `Assignments` section, go to the `Midterm` assignment and upload your completed files. If you have trouble submitting the files to Canvas, email them to lzhao@wharton.upenn.edu and rosesamk@sas.upenn.edu.

The submission folder will be closed sharp at 9:10PM.

On Site Help:

We will answer any clarification questions. We may also help out with some minor code issues. We will, however, not provide any answers as to what functions to use for example.

Raise your hand if you want to talk to one of us.

In case of emergency, here is Linda's cell: 6106590187 (text or call her)

Background

The outbreak of the novel Corona virus disease 2019 (COVID-19) was declared a public health emergency of international concern by the World Health Organization (WHO) on January 30, 2020. Upwards of 112 million cases have been confirmed worldwide, with nearly 2.5 million associated deaths. Within the US alone, there have been over 500,000 deaths and upwards of 28 million cases reported. Governments around the world have implemented and suggested a number of policies to lessen the spread of the pandemic, including mask-wearing requirements, travel restrictions, business and school closures, and even stay-at-home orders. The global pandemic has impacted the lives of individuals in countless ways, and though many countries have begun vaccinating individuals, the long-term impact of the virus remains unclear.

The impact of COVID-19 on a given segment of the population appears to vary drastically based on the socioeconomic characteristics of the segment. In particular, differing rates of infection and fatalities have been

reported among different racial groups, age groups, and socioeconomic groups. One of the most important metrics for determining the impact of the pandemic is the death rate, which is the proportion of people within the total population that die due to the the disease.

There are two main goals for this case study.

1. Number of deaths vary drastically across State. We want to find out how State relate to the death rate.
2. There have been studies on COVID racial disparities. Is there evidence in our data to show that the proportion of race relates to the death at county level?

To make our case study here simple and manageable in a timely fashion, we have assembled a subset of data called: `covid_county_midterm_spring_2022.csv`. It includes county level total number of deaths by a chosen date, together with selected demographic information. `State` names is also included in the data. The name of each variable should be self-explanatory.

1. Data preparation

In this case study, we have created `covid_county_midterm_spring_2022.csv` based on the following two cleaned datasets. We will focus on the east coast, i.e., Connecticut, District of Columbia, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Maryland, Rhode Island and Vermont.

- **covid_county.csv**: County-level socialeconomic information that combines the above-mentioned 4 datasets: Income (Poverty level and household income), Jobs (Employment type, rate, and change), People (Population size, density, education level, race, age, household size, and migration rates), County Classifications
- **covid_rates.csv**: Daily cumulative numbers on infection and fatality for each county

Among all data, the unique identifier of county is `FIPS`.

Question 1: (3 parts)

What is a good way to measure COVID death rate? There are quite a number of counties with a very low or zero number of deaths. We have proposed an effective way of handling such an imbalanced situation. In the following chunk, we created a new measurement of death-rate then applied the log function to get the variable labeled as `log_death_rate`. We then created the data `covid_county_midterm_spring_2022.csv` by combining the `log_death_rate` with a subset of county level demographic information. The process of creating this data is shown in the following R-chunk, labeled as `data prep`. Read through the `data prep` chunk carefully and please answer the following questions.

Note:

- Do not run this chunk!!!!! Notice the `eval=F`.
- Regardless your answers to the following questions, it will not affect the remaining case study at all.

i. The total number of death for each county is gathered by which day?

Answer:

ii. Use plain language to describe how the death-rate (before taking log) is defined for each county?

Answer:

```
# This is how we created the covid_county_midterm_spring_2022.csv
# DONOT RUN this chunk

# county-level socialeconomic information
county_data <- fread("data/covid_county.csv")
# county-level COVID case and death
```

```r
covid_rate <- fread("data/covid_rates.csv")

# northeast regions
northeast <- c("CT", "DC", "ME", "MA", "NH",
               "NJ", "NY", "PA", "MD", "RI", "VT")

covid_county_temp <- covid_rate %>%
  filter(date == "2020-09-01") %>%
  mutate(log_death_rate = log( (cum_deaths+1)/(TotalPopEst2019+2) )) %>%
  select(FIPS, cum_deaths, log_death_rate)

# join with county-level demographic data
covid_county_temp <-
  left_join(covid_county_temp,
            county_data,
            by = "FIPS") %>%
  filter(State %in% northeast) %>%

  drop_na()

# take a subset of the demographic info
covid_county_sub <- covid_county_temp %>%
  select(log_death_rate, State, Deep_Pov_All, PovertyAllAgesPct, PerCapitaInc, UnempRate2019, PctEmpFIR

# output
fwrite(covid_county_sub, "data/covid_county_midterm_spring_2022.csv")
```

We next read the pre-processed data `covid_county_midterm_spring_2022.csv` into R and label it as `covid_county`.

Note the `covid_county_midterm_spring_2022.csv` is stored in a directory called `data`.

`covid_county` will be used throughout the midterm.

```r
covid_county <- read.csv("data/covid_county_midterm_spring_2022.csv") # let's not use fread to avoid un
```

iii. For the `covid_county`, how many variables are included here? How many counties are there in this data? Are there any missing values in this data?

Answer:

2. EDA

During the course of pandemic, policies are usually implemented at state level and thus vary among states, which may further lead to the variability of death rate among states. We first study the death rates at state level via the following EDA.

Question 2: (4 parts)

i. Report number of counties by State.

Answer:

ii. Create the median `log_death_rate` by State and output the the table. Based on the table which state has the highest median of `log_death_rate`? And what is the corresponding highest median of `log_death_rate`. **No need to use R to produce the answers here.**

Answer:

iii. To compare `log_death_rate` across `State` and the county level variability within `State`, make a box-plot of `log_death_rate` by `State`. Use no more than two sentences to describe the comparison of `log_death_rate` by `State` and the variability within each `State`. Why does DC not have a box in the plot? (No need to order the boxes by median.)

Answer:

iv. We will use the proportion of white non-hispanic `WhiteNonHispanicPct2010` to explore how race is related to the death rate. Plot a scatter plot of `log_death_rate` vs `WhiteNonHispanicPct2010`. Use one sentence to summarize the trend of `log_death_rate` vs `WhiteNonHispanicPct2010` based on the scatter plot.

Answer:

3. Analyses

There are a number of studies indicating that COVID affected minority groups more. In the following analyses, we focus on the effect of `WhiteNonHispanicPct2010` over `log_death_rate`.

Question 3: (1 part)

3.1 fit1: `log_death_rate` vs. `WhiteNonHispanicPct2010` controlling `State`

Run a regression of `log_death_rate` vs. `WhiteNonHispanicPct2010` controlling `State` (without interactions) as `fit1`.

i. Report the summary table of `fit1`. Is `WhiteNonHispanicPct2010` a significant variable at the .01 level controlling for `State`? Use no more than 2 sentences to interpret the coefficient of `WhiteNonHispanicPct2010` over the `log_death_rate` in `fit1`.

Answer:

Question 4: (9 parts)

3.2 fit.final

In this section, using all possible variables available in `covid_county`, we will build a final parsimonious model to identify a set of important variables that are related to the `log_death_rate`.

As you have seen `State` explains a large portion of variability in `log_death_rate`, **we will keep `State` in all the following questions.**

i. Use LASSO to pick up a few variables in addition to `State`. To be specific let us control the following settings to get the same results.

- Use `set.seed(1)` to control the cross-validation
- Use 12-fold cross validations
- Force `State` in all the LASSO models
- Pick up the variables using `lambda.1se`

Report the variables selected by the LASSO.

Answer:

ii. Is `lambda.1se` a reasonable choice to have a small testing error? Use the LASSO plot to support your statement.

Answer:

In case you can't get LASSO to work by forcing `State` in the models you may run LASSO without forcing `State` in the models. But do include `State` in the remaining questions regardless

iii. Start with the set of variables obtained from your LASSO output. Run backward elimination until all variables are significant at 0.01 level as the final set of variables. Always keep `State` in the model. Show the backward selection procedure and report the final set of variables. (Hint: `regsubsets()` is not applicable here).

Answer:

iv. Run a final model `fit.final` of `log_death_rate` vs the set of variables from backward elimination (Q4.iii). Also include `Age65AndOlderPct2010` regardless since we know the death rate among the elderly is higher. Report the summary of `fit.final`.

In case you can't get LASSO to work use the following set of variables to start your `fit.final`: State, PctEmpServices, PopDensity2010, Age65AndOlderPct2010, WhiteNonHispanicPct2010, HiCreativeClass2000. Note: this is unnecessarily the LASSO output.)

Answer:

v. Is `WhiteNonHispanicPct2010` significant at .01 level controlling for all other variables in `fit.final`?

Answer:

vi. Is `State` significant at .01 level controlling for all other variables in `fit.final`? Which State has the largest `log_death_rate` controlling for all other variables in `fit.final`? (Hint: you can answer the question directly from the summary table of `fit.final`. No need to use R.)

Answer:

vii. Are the linear model assumptions reasonably met in `fit.final`? Provide residual and normal plots for `fit.final` and summarize your model diagnoses. (No more than 3 sentences).

Answer:

viii. Based on `fit.final`, write down the **prediction equation** for a county with the following characteristics:

- State: NJ
- PctEmpServices: 48.5
- PctEmpFIRE: 11.3
- PctEmpTrans: 7.97
- PctEmpMining: 0.0114
- PopDensity2010: 13731
- TotalPopEst2019: 672391
- ForeignBornPct: 42.77
- Age65AndOlderPct2010: 10.4
- WhiteNonHispanicPct2010: 30.8
- UrbanInfluenceCode2013: 1
- HiCreativeClass2000: 1
- Deep_Pov_All: 7.15
- NetMigrationRate1019: -3.255

No calculation needed.

Answer:

ix. Assume all linear model assumptions are met. Write a brief summary of your findings based on `fit.final`. (No more than 4 sentences).

Answer:

End of the case study!!!!

Appendix

Data Summary

The data comes from several different sources:

1. County-level infection and fatality data - This dataset gives daily cumulative numbers on infection and fatality for each county.
   - NYC data
2. County-level socioeconomic data - The following are the four relevant datasets from this site.
   i. Income - Poverty level and household income.
   ii. Jobs - Employment type, rate, and change.
   iii. People - Population size, density, education level, race, age, household size, and migration rates.
   iv. County Classifications - Type of county (rural or urban on a rural-urban continuum scale).