COVID-19 Case Study Midterm

your name

6:00-9:00PM (ET), 03/29/2021

Instruction

We have designed a two hour exam as planned. Due to virtual environment, we will allow anyone to stay for a three hour period. All the teaching team will be available from 6:00 - 9:00 PM. The submission will be closed sharp at 9:00PM.

Midterm Zoom link: https://upenn.zoom.us/j/92401814411?pwd=Q1E3cnBWOWpHNE5zelZXZWRCRHVJQT09

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 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: 1 part**
- **Question 3: 5 parts**
- **Question 4: 5 parts**

Data needed for the Midterm: `/canvas/Files/Exams/Midterm/Midterm Spring 2021/data/covid_county_midterm.c`

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 to submit the files to canvas email them to lzhao@wharton.upenn.edu and junhui@wharton.upenn.edu.

The submission folder will be closed sharp at 9:00PM.

On Site Help:

- Any clarification questions should be posted in the chat to the class.
- If you want to talk to one of us you may do the following (for example, if you are stuck somewhere. . . .We will try to debug for you.)
  - 5-6 break-out rooms are created
  - **Raise your hand** (We will send you to an available break-out room one at a time.)
  - Send a private chat to one of us
  - Email to the entire team
- Our emails:

lzhao@wharton.upenn.edu junhui@wharton.upenn.edu
cyfang@wharton.upenn.edu
farnik@sas.upenn.edu

niparkes@wharton.upenn.edu
rosesamk@sas.upenn.edu

Linda's cell: 6106590187

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. Covid seems to target elder people's lives. Is there evidence in our data to show that proportion of elder people indeed relates to the death at county level.

1. Data preparation

To make our case study here simple and manageable in a time fashion, we have assembled a subset of data called: `covid_county_midterm.csv`. It collects county level death rate, labeled as `log_death_rate`, as well as a subset of demographic information based on the two cleaned data:

- **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

Death rate

What is a good way to measure Covid death rate? There are quite a number of counties with a very low or none number of deaths. We have created a new measurement of death rate as follows:

- The total number of death for each county is gathered by November 1st, 2020
- The death rate is calculated as the total number of deaths plus 1 in the county divided by the total population of the county plus 2. We then apply the function log to get the `log_death_rate`.

Read data

We are ready to read the data `covid_county_midterm.csv` into R.

To simplify the analyses further, we created a subset here called `covid_county_sub` for us to use in the entire case study.

```
covid_county <- fread("data/covid_county_midterm.csv")
# covid_county <- fread("covid_county_midterm.csv")

covid_county_sub <- covid_county %>%
  select(log_death_rate, State, Deep_Pov_All, PovertyAllAgesPct, PerCapitaInc, UnempRate2019, PctEmpFIR
```

2. EDA

During the course of pandemic, we have witnessed that many policies are carried out at state level. For example when to reopen after the March lock-down. To see the variability of death rates among states we suggest you to go through the following EDA process.

Question 1: (3 parts)

1) Create the average `log_death_rate` by State. Show the histogram of the average `log_death_rate` by State. Use no more than three sentences to summarize the variability of the average `log_death_rate` by State.

2) To see within `State` county level variability, make box-plots `log_death_rate` by `State`. Use no more than two lines to describe the variability of `log_death_rate` by `State`.

3) What is the state with the highest average `log_death_rate`, and what is this rate?

3. Analyses

In the following analyses, we try to find out factors related to `log_death_rate`.

3.1 fit1: `Age` and `log_death_rate`

There are a number of studies indicating covid claimed most of elder lives. Let us start with a simple regression of `log_death_rate` vs. `Age65AndOlderPct2010`.

Question 2: (1 part)

1) Is `Age65AndOlderPct2010` a significant variable at .01 level in this analysis? Show the p-value.

3.2 fit2: `Age` and `log_death_rate` controlling for `State`

Question 3: (5 parts)

How do `Age65AndOlderPct2010` and `State` collectively affect `log_death_rate`. In `fit2`, run a linear model of `log_death_rate` vs `State` and `Age65AndOlderPct2010` (without interactions).

1) Is `Age65AndOlderPct2010` significant at .01 level in this model?

2) How to interpret the coefficient of `Age65AndOlderPct2010` over the `log_death_rate` in `fit2`? How would you explain the difference in effects of `Age65AndOlderPct2010` in `fit1` vs `fit2`. (No more than 3 sentences).

3) Perform a test to see if `State` is significant in this model at .01 level?

4) Based on `fit2`, what is the estimated `log_death_rate` for a county in `NJ` and `AL` given `Age65AndOlderPct2010 = 20` respectively. Note that the base level of `State` is `AL`. **Show your formula and evaluate the final values. Do not use predict() in this question.**

a) For NJ: `log_death_rate = `.

a) For AL: `log_death_rate = `.

5) Are the linear model assumptions reasonably met in `fit2`? Provide residual and normal plots for `fit2`. Use no more than three sentences summarizing your model diagnoses.

3.3 fit.final

In this section, using all possible variables available in `covid_county_sub`, we will build a final parsimonious model and to identify a set of important variables that are related to the `log_death_rate`. We will not fine-tune the final model.

As you have seen `State` effect explains a large portion of variability in `log_death_rate`, we will lock `State` in all the analyses.

Question 4: (5 parts)

Important Remark: You are going to run LASSO to pick up a few variables in addition to the `State`. In case you can't get LASSO to work go to **2)** directly and use the following set of variables to get your `fit.final`: State, PctEmpServices, PopDensity2010, Age65AndOlderPct2010, WhiteNonHispanicPct2010, HiCreativeClass2000. (Note: this is not necessarily the LASSO output.)

1) Use LASSO to pick up a few variables in addition to `State`. List variables output from the above LASSO.

To be specific let us control the following settings to get the same results.

- Use set.seed(1) to control the cross-validation errors.
- Use 10-fold cross validations.
- Force `State` in all the LASSO models.
- Pick up the final set of variables using `lambda.1se`

2) Run a final model `fit.final` of `log_death_rate` vs `State` and the set of variables obtained from your LASSO output. Also include `Age65AndOlderPct2010` regardless whether it is in your LASSO output or not. (You can easily specify the lm() variables without any algorithm.)

Report the summary of `fit.final`.

3) Is `State` significant at .01 level in this model? Is `Age65AndOlderPct2010` significant at .01 level in this model?

4) Controlling for all other variables in `fit.final`, which state has the highest `log_death_rate` and what is the value?

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

End of the case study!!!!