Group exercise 3: The spread of COVID-19 in the US

DATA5207: Data Analysis in the Social Sciences

Dr Shaun Ratcliff

In the final two labs of the unit, we will be working on your last assessable group exercise.

This task, which counts towards your group work grade, should take you no more than approximately four hours for each group member to complete over the next two sessions. It is worth 10 per cent of your total grade for the unit and will be conducted in your groups for the assessable group projects. You should submit your final written document as a *Markdown* file through a link that will be provided on canvas. You do not need to knit this, just submit the RMD with the files needed to make it run in a zipped folder. Only one member of your group needs to submit your work. As long as the groups are properly registered on canvas, all members of the group will receive the full grades awarded to the group.

The project

We will finish the unit by working on what I think is an interesting and contemporary problem. I am sharing with you some data I have been working with to predict the extent of COVID-19 confirmed cases in the United States. Using these data and the methods we have covered so far, you will **write a short article in the style of a data journalist** that explains the spread of COVID-19 in the United States. This should be written in a clear and accessible manner and the explanations should include high quality visualisations. In preparation for this, you should explore your data and conduct descriptive analysis. You can run a regression model, although this is optional. Remember, the key goal is to convey your findings in a clear and accessible way.

The teaching team will be available during the labs and to talk during their office hours, to provide any help you need.

You will do this using three datasets. One contains a time series of COVID-19 cases (which you will use to create your dependent variable). The other two contain data that can be used for potential predictors: Google mobility data by county (also a time series) and demographic and other information on individual counties (static across time).

I have provided you with a cleaned and combined version of these data. This is the file *covid.data_cleaned.Rdata*. This does not include all of the variables from the original data, though. The code used to download and clean the data is shown below.

Start with the cleaned file, but add variables from these datasets as needed. You will need to really get into the data to understand it and then think about how you are going to use it to answer your question. The county data (which contains most of your potential predictors) is quite large. If you dig into this for additional predictors, you might want to use the `select()` function to only pull out those predictors you intend to use before combining the data.

Your data

You have three datasets for this task. Two time-series datasets: the first is the daily number of new cases of COVID-19 in each county in the United States from January 2020 to May 2022, and the other tracks daily

changes in mobility in each county from February 2020 to May 2022. You also have a set of stationary data on other features for each county, including demographic, economic and climate characteristics.

The first of these datasets is the file containing information on new confirmed COVID-19 cases in the US, by county. These data are sourced from the Johns Hopkins dataset, available here.

I have downloaded and cleaned these data so you do not have to. I have used them to create a file that contains the number of confirmed cases and a 14-day rolling average for each county in the US. The file is called `covid.RData` and it is saved in the data folder on the canvas page for this session.

So you can see how I did this, the code is included here:

```
# load data

confirmed.cases.data <- read.csv("https://raw.githubusercontent.com/
                                  CSSEGISandData/COVID-19/master/csse_covid_19_data/
                                  csse_covid_19_time_series/
                                  time_series_covid19_confirmed_US.csv")

# load packages

library(tidyr)
library(DataCombine)


# clean data

confirmed.cases.data2 <- confirmed.cases.data %>%
  dplyr::rename(state = Province_State,
                county = Admin2) %>%
  dplyr::select(-UID, -iso2, -iso3, -code3,
                -Lat, -Long_, -Combined_Key, -Country_Region) %>%
    gather(date, confirmed.cases,-state,-county,-FIPS) %>%
  dplyr::mutate(date = gsub('X', '', date),
                date = as.Date(as.character(date), "%m.%d.%y"),
                county_state = paste0(county, ', ', state)) %>%
  arrange(county_state, date) %>%
  dplyr::mutate(lag = slide(.,
                                  Var = 'confirmed.cases',
                                  NewVar = 'new',
                                  GroupVar = 'county_state',
                                  slideBy = -1)[,'new'],
                new.cases = confirmed.cases - lag)


## calculate rolling averages and remove non-states

covid.smooth.data_county <- confirmed.cases.data2 %>%
    dplyr::group_by(county_state, date) %>%
    dplyr::summarise(new.cases = sum(new.cases)) %>%
      dplyr::group_by(county_state) %>%
    dplyr::mutate(cases_14days = zoo::rollmean(new.cases,
                                               k = 14, fill = 0)) %>%
  dplyr::ungroup() %>%
```

```
  mutate() %>%
  merge(confirmed.cases.data2 %>%
          dplyr::select(county_state, date, county, state)) %>%
  filter(!state %in% c('American Samoa',
                       'Diamond Princess',
                       'Grand Princess',
                       'Guam',
                       'Northern Mariana Islands',
                       'Puerto Rico',
                       'Virgin Islands'))
```

This provides us with our dependent variable, the confirmed cases column (and also the rolling 14-day average).

The next dataset is Google mobility data. This shows how mobility patterns have changed in different US counties, and can be obtained with the code:

```
google.mobility <- read.csv('https://www.gstatic.com/covid19/mobility/
                            Global_Mobility_Report.csv') %>%
  filter(country_region == 'United States')
```

This is a large file and many people struggle to download it, so I have saved the csv file in the Data folder in the canvas module for this lab.

The documentation for this can be found here. The county-level variable is called `sub_region_2` in this dataset.

The rest of your predictors can be found in this file here:

```
county.data <- read.csv('https://raw.githubusercontent.com/JieYingWu/
COVID-19_US_County-level_Summaries/master/data/counties.csv')
```

The codebook for these data can be found here. Additional information on these data are available here.

The cleaned file

I conducted some additional cleaning and rolled these into the file *covid.data_cleaned.Rdata*. There is a lot here and you can just run your analysis on these data if you want. Or you can use the syntax below to add additional variables to the dataset and your analysis.

The cleaned file contains:

COVID-19 data

You can use any of these variables as your dependent variable in your analysis, or transform them as you wish.

- `new.cases` - The number of new daily cases
- `cases_14days` - The rolling 14 day average of new cases
- `new.cases.lag` - A one-day lag of new cases

Date and location data

- `date` - The date of the observation
- `county` - The county in which the observation was taken
- `state` - The state in which the observation was taken
- `county_state` - The combined county and state in which the observation was taken

Demographic data (this is stationary and does not change by date)

- `total.population` - The total population of the county
- `pop.density` - The population density of the county (per square mile)
- `housing.density` - The density (per square mile) of housing units
- `age.65.plus` - The share of the population that is aged 65 and older
- `pcnt.university` - The share of the population with a university education
- `pcnt.less.than.high.school` - The share of the population that did not complete high school or obtain further education
- `avg.hhold.size` - The mean size of households in the county
- `transit.scores` - How well a county is served by public transit
- `median.income` - Estimated median household income in 2018

Climate data (this is stationary and does not change by date)

I think these are self explanatory.

- `avg.dec.tem`
- `avg.jan.temp`
- `avg.feb.temp`
- `avg.winter.temp`

Healthcare resource data (this is stationary and does not change by date)

- `icu.beds` - Number of ICU beds per county
- `active.physicians` - Active Physicians per 100,000 Population, 2018
- `active.patient.care.physicians` - Total Active Patient Care Physicians per 100,000 Population, 2018

Mobility data

- `mobility.transit` - Mobility trends for places like public transport hubs such as subway, bus, and train stations. Counties with no public transport have been scored zero
- `mobility.workplaces` - Mobility trends for places of work
- `workplaces.lag14` - A 14 day lag for this
- `mobility.grocery` - Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies

How these data were cleaned and combined

Here is the code used to clean and combine these data:

```
# patterns for cleaning county names

patterns <- c(' Borough| Census Area| County| Parish| city')


# covid case data

covid.data <- covid.smooth.data_county %>%

  # county data

left_join(county.data %>%

            # grab file with full state names

  left_join(read.csv('Data/fips concordance.csv') %>%
              dplyr::select(State = Alpha.code,
                            state_full = Name)) %>%
        dplyr::mutate(county = gsub(patterns, '', Area_Name),
                      county_state = paste0(county, ', ', state_full)) %>%

    dplyr::select(county_state,
                  pop.density = Density.per.square.mile.of.land.area...Population,
                  age.65.plus = Total_age65plus,
                  total.population = POP_ESTIMATE_2018,
                  median.income = MEDHHINC_2018,
                  avg.hhold.size = Total.households..Average.household.size,
                  avg.dec.temp = Dec.Temp.AVG...F,
                  avg.jan.temp = Jan.Temp.AVG...F,
                  avg.feb.temp = Feb.Temp.AVG...F,
                  icu.beds = ICU.Beds,
                  active.physicians = Active.Physicians.per.100000.Population.2018..AAMC.,
                  active.patient.care.physicians = Total.Active.Patient.Care.Physicians.per.100000.Popul
                  pcnt.university = Percent.of.adults.with.a.bachelor.s.degree.or.higher.2014.18,
                  pcnt.less.than.high.school = Percent.of.adults.with.less.than.a.high.school.diploma.20
                  housing.density = Density.per.square.mile.of.land.area...Housing.units,
                  transit.scores = transit_scores...population.weighted.averages.aggregated.from.town.ci


  merge(google.mobility %>%
          dplyr::mutate(county = gsub(patterns, '', sub_region_2),
                        county_state = paste0(county, ', ', sub_region_1),
                        date = as.Date(date),
                        mobility.transit = ifelse(is.na(transit_stations_percent_change_from_baseline),
                                  transit_stations_percent_change_from_baseline)) %>%

  dplyr::select(county_state,
                date,
                mobility.transit,
```

```
                mobility.workplaces = workplaces_percent_change_from_baseline,
                mobility.grocery = grocery_and_pharmacy_percent_change_from_baseline)) %>%

  # convert to rates

  mutate(new.cases = new.cases / (total.population / 100000),
         icu.beds = icu.beds / (total.population / 100000),
         age.65.plus = age.65.plus / total.population,
         avg.winter.temp = (avg.dec.temp + avg.jan.temp + avg.feb.temp) / 3)


# lag new cases

new.cases.rate.lagged <- DataCombine::slide(covid.data,
                                      Var = 'new.cases',
                                      NewVar = 'new.cases.lag',
                                      GroupVar = 'county_state',
                                      slideBy = -1)

# lag workplace mobility

workplace.lag14 <- DataCombine::slide(covid.data,
                             Var = 'mobility.workplaces',
                             NewVar = 'workplaces.lag14',
                             GroupVar = 'county_state',
                             slideBy = -14)


# combine

covid.data <- covid.data %>%

  merge(new.cases.rate.lagged %>%
           dplyr::select(county_state, date,
                         new.cases.lag)) %>%

    merge(workplace.lag14 %>%
           dplyr::select(county_state, date,
                         workplaces.lag14))
```

Your task

Develop a data journalism project explaining the spread of COVID-19 in the US. Why has it impacted some areas more than others? This should be accessible, with visualisations, and make the significance of the topic and findings clear to the reader.

1. Start by developing some ideas on what factors may influence the spread of COVID-19. Write these ideas down, and use them to select several predictors from these data (my suggestion would be approximately five, but more is fine).
2. Grab additional variables from datasets as needed so that you have a single file with all the variables you need. You are to use the COVID-19 case data as the dependent variable. Source your predictors from the `county.data` file and the Google mobility data.

3. Look at some of the descriptive results for your chosen variables, including their distributions as well as their relationship to the confirmed cases variable. Plot these.

4. Once you have done this, you can fit a regression model to these predictors, using the confirmed cases variable from the first dataset as your dependent variable. **This is optional, and is not required though.**

5. What do the results tell you? Write this down in a clear and accessible way. Explanations should include high quality visualisations. **This is the most important part of the assessment.**

You will have the entirety of the last two labs of the semester to work on this project.

Submit your work

Once you are finished, submit the RMD your are working on using the *Group project 3* assessment on canvas, *along with your data.* Only one version needs to be submitted. Full grades will be allocated to all registered members of the group.

All the files needed to run your code should be uploaded in a zipped folder. We should be able to run the code without changing it.

You will be marked on the quality of your R code (including whether it runs for us without errors), how well you have justified your variable selection, the proper use of appropriate methods, and the quality of your visualisations.

If you need help

If you have any questions, do not hesitate to ask us for help. We cannot do the work for you — this is an assessment – but we can provide some advice.

Good luck with the exercise!