Day 6 Group exercise: The spread of COVID-19 in the US

DATA5207: Data Analysis in the Social Sciences

Dr Shaun Ratcliff

In both labs today, we will be working on an assessable group exercise. You will have a day to finish and upload your work (see the *Group assignment 3* assessment on canvas).

This task, which counts towards your group work grade, should take you no more than three or four hours to complete (most of which will be time we give you to work on this in the labs). 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, which should have approximately four or five members. You should submit your final written document as a *Markdown* file through a link that will be provided in this module. You do not need to knit this, just submit the RMD. 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 assessable group task for today

We will finish the week 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 fit a model to county-level data from the US estimating the predictors for the spread of the coronavirus in the US.

For this exercise, you will work in your groups to solve the problem. I will move between the groups helping you as you go.

You will do this using three datasets, which you need to combine. 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).

Clean and combine these three 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. You might want to use the `select()` function to only pull out those predictors you intend to use before combining the data.

I have provided you with instructions for downloading three datasets, although you are expected to work out some of the data cleaning yourself (but I will be here to help you). We will download them all directly from the internet into *R*. If for some reason one of the following downloads don't work, just download the file directly and then run the rest of the code. If you do download the data directly, do not open the dataset in another program (like Excel or Numbers). Some spreadsheet programs have maximum cell numbers for files, and may trim excess rows or columns, losing you data.

Once you have done this, fit a regression model to your chosen predictors, using the confirmed cases variable from the first dataset as your dependent variable, and follow the rest of the instructions detailed below.

Your data

The first dataset we are going to download is a 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'))


save(covid.smooth.data_county, file = 'Data/Day 5/covid.RData')
```

This provides us with our dependent variable, the confirmed cases column (and also the rolling 14-day average). We obtain our predictors from two other datasets. The first of these 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?cachebust=722f3143b586a83f') %>%
  filter(country_region == 'United States')
```

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.

Your task

1. Start by developing a theory 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. Clean and combine these three 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. To get your predictors, you need to use the demographic data and Google mobility data.

You will need to use the county variable to merge the three datasets. Note that in its raw state, this variable has a different name in each dataset and may be formatted differently. Use the notes from Day 5 as a guide on how to deal with this.

You want to use all of the data from the time series available in the COVID-19 case and Google mobility data, and you can assume the demographic data stays constant over time (reasonable given the short time frame of these data), with the additional columns for each demographic predictor remaining constant for each date-county pair.

Your combined dataset should have a column for COVID-19 cases, column(s) for Google mobility predictors (if you choose to use them), and a set of columns for demographics (with data in these columns this same for each date for each county).

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, fit a regression model to these predictors, using the confirmed cases variable from the first dataset as your dependent variable.
5. What do the results tell you? Write this down.

You will have the entirety of both labs today to work on this project.

Submit your work

Once you are finished, submit the RMD your are working on using the *Group assignment 2* 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.

If you need help

If you have any questions, do not hesitate to ask us for help. During the labs, the teaching team will be available to talk you through the project on Zoom. We will be moving through the breakout groups and can also be called to assist you. Outside of the sessions today, you can also post questions on the Ed discussion board.

We cannot do the work for you — this is an assessment – but we provide some advice.

Good luck with the exercise!