Winter 2024
PPHA 39930
Instructor: Amir Jina
Individual assignment 2
Due: Wednesday 31st January, 2024

Directions: Please submit a PDF write-up with your answers to questions and any required figures. For data analysis, also submit the file containing your analysis. For example, submit the code / output of your coding language of choice (e.g., a .R file).

Estimating climate impacts

In this assignment you will *approximately* replicate a paper by Olivier Descheñes and Michael Greenstone entitled "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US". I'll refer to this as DG2011 from now. This paper looks at the relationship between mortality and temperature in the United States. It builds on an earlier paper from 2007 that examined the relationship between crop yields and temperatures, and is one of a handful of research articles that launched our new understanding of climate impacts upon society.

Begin by loading the dataset "`homework2_dg2011_rep_nomiss.csv`" into R (or some other language). This dataset has some important differences from the DG2011 data. First, for reasons of confidentiality, I have only provided the publicly accessible mortality data.[1] Second, it only contains data on deaths for those aged over 65 years. This means that you won't be able to exactly replicate DG2011, but you'll get close. Weather data is the same as used in DG2011. Temperatures are given as the count of the number of days where the average temperature falls within a 10°F range.[2] Descriptions of the variables in the dataset are below. For more details, see descriptions in DG2011.

Load the data and answers the questions that follow. Most will ask you only to interpret the output of the paper replication, not to be able to (for example) write out the complicated regression command yourself. The intention is to build intuition for how and why we can estimate causal effects of climate on human outcomes.

Variable name	**Explanation**
`county`	Name of county, State
`countycode`	5-digit county FIPS code
`statecode`	2-digit state FIPS code
`year`	Year when deaths and population were recorded
`deaths`	Number of deaths of over-65s
`population`	Population of over-65s
`cruderate`	Death rate (deaths per 100,000 population)
`tday_lt10`	Number of days with average temperature <10°F
`tday_X_Y`	Number of days with average temperature ≥X°F and <Y°F
`tday_gt90`	Number of days with average temperature >90°F
`cdd65`	Cooling degree days (reference temp 65°F)
`hdd65`	Heating degree days (reference temp 65°F)
`prec_X_Y`	Annual rainfall ≥X and <Y inches
`prec_gt60`	Annual rainfall ≥ 60 inches
`division`	Census division
`ssyy`	State×Year indicator

[1] This was obtained from the Center for Disease Control (CDC) Wonder platform. All counties with fewer than 10 deaths in a year are redacted. I've removed all counties that have missing data.

[2] Regrettably, all the data are in Fahrenheit because it's a paper about the United States!

Everything in the problem set can be calculated from the main dataset, plus the "homework2_county_avetemp.csv" file. For question 3, take the average through time within counties and merge in the "homework2_county_avetemp.csv" data. The merge won't be perfect, but you can drop unmerged counties.

1. First, let's look at temperatures across the US. Take a population-weighted average across all temperature variables for the entire country.

(a) Plot a histogram of number of days within each temperature range. This will be like figure 1 of DG2011, without the future projections.

(b) What is the average number of days above 90°F per year that is expected as a population-weighted average across the US?

(c) What county has the highest number of days above 90°F per year in the country? How many counties have, on average over the sample period (1968-2002) experienced zero days above 90°F per year?

2. Now, we'll try to get a sense of mortality rates. These are usually expressed as "deaths per 100,000 population". In our case, it's only for over-65s.

(a) Find out what the national all-age mortality rate is (via a google search!). How does the national average over-65 mortality rate in our data compare to the national all-age mortality rate that you just found?

(b) How many deaths from 1968-2002 are recorded in the data in total?

3. As mentioned in class, one way to think about temperature effects on health is to look at the *cross-section.* Let's look at the averages of counties' mortality and temperature through time. You can calculate those means by group, where the county is the group.

In addition to just average temperature, generate a variable "hotdays" that is equal to the sum of all days over 70°F, and another variable "hotterdays" that is equal to the sum of all days over 80°F. You should also have a variable "normal_1981_2010", which is the climatological temperature in each county.

(a) You first want to check the relationship between county average temperatures and over-65 mortality rates. Plot this figure and fit a regression line to the data. What is the slope? What pattern emerges?

(b) Now you think that maybe the extreme days have a more important effect than average temperature. Plot two more figures, one for mortality rates versus hot days, and another for mortality rates versus hotter days. Fit regression lines to each. What pattern emerges?

(c) Try to explain why you might observe this pattern in a few sentences.

4. You conclude that the time-dimension of the data will be useful. Reload the full dataset. Start by examining only a subset of counties through time. For this question, you should drop counties or subset counties so that your dataset contains *only*:

- Mobile County, Alabama
- Cook County, Illinois
- Los Angeles County, California
- Miami-Dade County, Florida

(a) Plot a scatterplot of the death rates in these counties versus the hotter days variable. Plot a line of best fit through all the data points. Describe the relationship that you would conclude from looking at this plot.

(b) Now plot a scatter of them again, but this time only fit a line to each county individually, not to all counties. What do you notice about the slopes of the four lines you have plotted? How does it compare to part (a)? Why might you be more confident or less confident in the colclusion you draw from this plot than from the plot in part (a)?

5. Finally, you have a sense that it's the *within county* variation that you care about. This is where you'll fully try to replicate DG2011. The regression equation in the paper is as follows:

$$\underbrace{Y_{ct}}_{\substack{\text{mortality rate in}\\ \text{county, c, in year, t}}} = \underbrace{\sum_j \theta_j^{TEMP} NumDays_{ctj} +}_{\substack{\text{represents nine separate}\\ \text{terms counting number}\\ \text{of days in temperature range}}} \underbrace{\sum_l \delta_l^{PREC} PREC_{ctl}}_{\substack{\text{represents 11 separate}\\ \text{terms indicating that}\\ \text{precip within certain range}}} + \underbrace{\alpha_c}_{\substack{\text{county-level}\\ \text{fixed effects}}} + \underbrace{\gamma_{st}}_{\substack{\text{state}\times\text{year}\\ \text{fixed effects}}} + \underbrace{\varepsilon_{ct}}_{\substack{\text{error /}\\ \text{residual}}}$$

One way to write this command in R to run this regression is written out below. There are others - feel free to choose your own. Notice a couple of things:

- One of each of the temp. and precip. variables is omitted. The omitted category acts as a reference category. i.e., the coefficient on 90F days compares the effect on a 90F day to the effect on the (omitted) 60-70F days
- The regression includes both county and state-year fixed effects
- The regression is weighted by population. This changes the interpretation to the effect on a randomly drawn person in the US, rather than the effect in a randomly drawn county in the US
- There's likely to be correlations across space and through time, so we "cluster" the standard errors at county level. This just means that our analysis allows county records to be correlated through time.

```
install.packages("lfe")
#
library("lfe")
#
dg2011 <- felm(cruderate
               ~ tday_lt10 + tday_10_20 + tday_20_30 + tday_30_40 + tday_40_50
               + tday_50_60 + tday_70_80 + tday_80_90 + tday_gt90
               # ^ temperature variables
               + prec_10_15 + prec_15_20 + prec_20_25 + prec_25_30 + prec_30_35
               + prec_35_40 + prec_40_45 + prec_45_50 + prec_50_55 + prec_55_60 +
                 prec_gt60
               # ^ precipitation variables
               | countycode + ssyy| 0 | countycode,
               # ^ fixed effects and clustering
               data=mortality_temp, weight=mortality_temp$population)
#
summary(dg2011)
```

(a) Using the full data again, run this regression. If you can, construct a figure like figure 2 in DG2011. If not, report the values that you estimate. Describe what these results mean for the relationship between mortality and temperature.

(b) Read the description of the regression on p164-165 of DG2011. What role do these fixed effects (α_c and γ_{st}) play in the analysis? They give two examples of things that the fixed effects control for. Think of two other examples for each of the two fixed effects in the analysis.

(c) Look at the coefficient on the hottest (>90°F) day term. You can interpret this as the change in the mortality rate for over-65s caused by replacing a single average 60-70°F degree day with a 90°F one. What percent change on the baseline mortality rate does this imply? Does this seem small or large to you?

6. Propose an extension of this analysis that might help you to understand how people are adapting. You can state this in words, and mostly rely on intuition (i.e., there is no need to write out an equation or formally test anything). What pattern would you expect to see and why?

7. Finally, think through some of the implications of the work you have just done. If you had previously only been able to access data on average temperature and average mortality rates (or just temperatures and mortality in a singe year), what would you have concluded about the relationship? How is your understanding changed by the results in question 5? What are the implications of the results in question 5 as the US warms under climate change?