Applied Economics Research using R: Session 1 - announcement

Seunghyun Lee (arslee@ucdavis.edu)

9/25/2021

Contents

Hi All,

I hope you had a great summer. By now, I expect that research is a top priority for many of you. I will be sharing some practical tips about coding and R in two sessions. In the first session, I will talk about efficient workflow using R for applied economics research. In the second session, I will walk through how to use R for geospatial data analysis. Two hours are too short to cover many things. But, I will try to make the materials generalizable enough so that you can easily apply the skills and knowledge you learn in my sessions to your research.

The main goal of my sessions is to give you a chance to look at what's behind the curtain in academic papers (i.e., things that papers don't tell you in detail but you need to do when you research).

Researchers tend to spend a lot of time 1) accessing, cleaning, exploring, and checking data and 2) improving efficiency, reproducibility, and readability of their codes. I mean really a lot. Although I still have many things to learn and improve, I would like to spend some time sharing things that I wish I had known in my second year.

In the first session, we are going to run the following panel regression of crop yields for corn and soybeans in Iowa, Illinois and Indiana over the period 1981-2019:

$$log(y_{cit}) = \beta_{c1} prec_{it} + \beta_{c2} prec_{it}^2 + \beta_{c3} gdd_{it} + \beta_{c4} hdd_{it} + \alpha_i + f_i(t) + \epsilon_{cit},$$

where y_{cit} denotes the yield for crop c in county i in year t. *prec* denotes precipitation, and *gdd* represents beneficial heat exposure (Growing Degree Days: degree days between 10C and 30C) and *hdd* represents harmful heat exposure (Heating Degree Days: degree days above 30C). We include county fixed effects α_i and county-specific time trend $f_i(t)$. The growing season for corn and soybeans in the three I states is approximately from April to September.

This crop yield model has become popular in the literature since the seminar work by Schlenker and Roberts (2009). You will learn more about the methodology and other related ones later in the course in more detail.

I do encourage you to spend some time trying to implement the following steps, although I will walk through them in the session

It is totally ok even if you do not make much progress. But, I would like you to at least try each step so that you can make most out of my session.

1. Download annual county-level yield and acreage data for corn (for grain) and soybeans for all counties in the I states from the USDA NASS quickstats for the period 1981-2019. (Note: There are multiple ways of doing this. I would like your workflow to be reproducible and automated, meaning that executing your script downloads all data you need at once. Please feel free to refer to Accessing Ag Data Using R)

2. Download weather data by clicking `Download All Monthly Data` from US County Weather once the app is fully loaded. (In a few weeks later, we are going to learn how to construct this data using the gridded daily temperature data from the PRISM Climate Group, county boundaries, and crop frequency map)

3. Using the data you downloaded, create a crop-county-year level panel dataset that contains columns of yield and weather variables necessary for your regressions. This will involve some data cleaning. (Tip: You can construct annual *gdd* or *hdd* by summing monthly degree days over the growing season (April to September))

4. Before running regressions, do some data exploration to check your data. You can compare your data with US Crops.

5. Run regressions for the following combinations.

    - crop : 1) corn, 2) soybeans
    - period : 1) full (1981-2019), 2) pre 2000 (1981-2000), 3) post 2000 (2001-2019),
    - regression weights : 1) no weight, 2) acreage
    - cluster standard errors by: 1) year, 2) state
    - time trend : 1) county-specific linear, 2) state-specific linear, 3) county-specific quadratic, 4) state-specific quadratic
    - weather variables : 1) (with precipitation) $prec$, $prec^2$, gdd and hdd, 2) (without precipitation) gdd and hdd

That is, you have $2 \times 3 \times 2 \times 2 \times 4 \times 2 = 192$ regression results.

6. Let's focus on β_4.

    - Obtain $\hat{\beta}_4$s and calculate their confidence intervals.
    - Compare $\hat{\beta}_4$s and their confidence intervals by changing one dimension at a time. (i.e., compare $\hat{\beta}_4$s for different crops 1) corn and 2) soybeans while keeping 1) for all the other dimensions).
    - Do you find any systematic differences between models in any dimensions? If so, what are they? Are they consistent with your intuition?

Lastly, **please feel free to reach out to me** if you have any questions before the session. I will be also happy to hear about your past experience in R, interests, or any concerns related to coding or R.