ARE 261: Problem Set #1

Due October 3, 2022

September 15, 2023

1 Temperature and Economic Outcomes

Part 1 of this problem set is designed to explore the relationship between temperature and economic outcomes. We are going to walk through the problem in various steps.

1. First, we are going to take a gridded dataset on temperatures for a single county and aggregate this to the county-year level.

2. Once we have successfully created our temperature aggregates, we are then going to explore the relationship between temperature variation and economic outcomes using county-year data from the entire United States.

1.1 Temperature Aggregation

Download the dataset: fips1001.dta. This is a gridded dataset from PRISM for Autauga County, Alabama. This data contains minimum and maximum temperature and precipitation from 1950-2012 on a 2.5km by 2.5km basis.

1. Use this data to construct four sets of temperature response variables:

   (a) A degree day measure using a sinusidal curve to model diurnal temperature cycle for a given day. See e.g., Schlenker and Roberts (2009). Create 3 separate degree-day measures, calculating degree days above 30, 32, and 34C

   (b) Binned temperature variables spanning 4 degree C bins: Below 0, 0-4, 4-8, 8-12, 12-16, 16-20, 20-24, 24-28, 28-32, Above 32

   (c) Restricted cubic splines with knots at 0 8 16 24 and 32

   (d) Piecewise linear functions, with breakpoints at 28 and 32.

2. Now, take the sum over these variables during the year, and then take the unweighted mean across grid cells within the county. Ideally, we would be weighting this collapse by population weights using Decennial Census Block population counts.

3. Check that you are able to replicate the variables for Autauga County (Fips 01001) in the following dataset: CountyAnnualTemperature1950to2012.dta.

1.2 US Climate Impacts: County-Year Damages

Download reis_combine.dta and merge it to CountyAnnualTemperature1950to2012.dta using county and year. The dataset reis_combine.dta contains information from the BEA's Local Area Personal Income and Employment: https://www.bea.gov/regional/downloadzip.cfm. The fips codes may not completely overlap between the datasets, because the BEA uses "county-equivalent" FIPS codes. Ignore the incomplete merge for the purposes of this assignment. Drop any remaining county-years that do not match between the datasets.

1. Explore the relationship between log tranformed emp_farm and the vector of binned temperature controls. Include additional controls for county FE, year FE. Provide an interpretation of the coefficient of the 32+ bin.

2. Explore the relationship between log tranformed inc_farm_prop_income/population (i.e. log(per capita farm prop income) using the restricted cubic spline. Include additional controls for county FE and year FE. Plot the predicted marginal effects with associated confidence intervals, and compare to the binned temperature response function above.

3. Use the binned temperature estimator to design a test for whether we observe treatment effect heterogeneity. One possibility for this test is to interact the temperature bins with the average # of days in a county for which the temperature falls into each respective bin (i.e. are counties that experience more 32+ days more/less responsive to temperature).

4. **[BONUS QUESTION]** The research community has a rudimentary understanding of treatment effect heterogeneity within each of these temperature response bins (aside from the heterogeneity in AC/mortality uncovered by Barreca et al. 2016). One interesting exercise (that may even be publishable) is the following:

    - In the presence of heterogeneous treatment effects, the OLS coefficient estimate is a weighted average of the heterogeneous treatment effects, where the weights depend on the relative size of the groups and the conditional variance within each group (See, e.g., Angrist and Krueger (1999); Wooldridge (2005); Angrist and Pischke (2009)). Derive the OLS regression weights for the 32C+ regressor using methods outlined in Angrist and Krueger (1999), Section 2.3. Replicate figure 3b in Angrist and Krueger, replacing schooling with the support of the 32C+ regressor. Where is most of the weight coming from within the 32C+ regressor? How does this affect our understanding of the the temperature-outcome dose-response relationship?

2 Hedonic Air Quality Analysis

This exercise examines the following research question: What is the relationship between changes in air pollution and housing prices? Please include a concise summary of your empirical results when appropriate. This is based on Ken Chay and Michael Greenstone (2004) "Does Air Quality Matter? Evidence from the Housing Market", and I think that reading the paper will help a lot with the analysis. The data used for this analysis (contained in poll7080.dta available on bCourses) is an extract from a combination of the 1972 and 1983 City and County Data Books, the EPA's Air Quality Subsystem data file, and the Code of Federal Regulations. The data is measured at the county-level in the United States.

2.1 Data Notes:

1. There are 1,013 observations and 26 variables at the U.S. county level. These are the counties with particulates pollution monitors both at the beginning and end of the 1970s and contain the vast majority of the U.S. population (over 80%).

2. The key variables are:

   dlhouse = change in log-housing values from 1970 to 1980 (1980 log-price minus 1970 log-price).

   dgtsp = change in the annual geometric mean of total suspended particulates pollution (TSPs) from 1969- 72 to 1977-80 (1977-80 TSPs minus 1969-72 TSPs).

   tsp7576 = indicator equal to one if the county was regulated by the Environmental Protection Agency (EPA) in either 1975 or 1976 and equal to zero, otherwise.

   mtspgm74 = annual geometric mean of TSPs in 1974.

3. The other relevant variables are:

   pop7080 = sum of 1970 and 1980 county populations; use to weight all regressions/analysis.

   ddens = 1970-80 change in population density.

   dmnfcg = change in % manufacturing employment.

   dwhite = change in fraction of population that is white.

   dfeml = change in fraction female,

   dage65 = change in fraction over 65 years old.

   dhs = change in fraction with at least a high school degree.

   dcoll = change in fraction with at least a college degree.

   durban = change in fraction living in urban area.

   dunemp = change in unemployment rate.

   dincome = change in income per-capita.

   dpoverty = change in poverty rate.

   dvacant = change in housing vacancy rate.

   downer = change in fraction of houses that are owner-occupied.

   dplumb = change in fraction of houses with plumbing.

   drevenue = change in government revenue per-capita.

   dtaxprop = change in property taxes per-capita,

   depend = change in general expenditures per-capita.

2.2 Questions:

Does Air Quality Get Capitalized into Housing Prices? The outcome of interest is the change in county housing prices during the 1970s. We want to estimate the "causal" effect of air pollution changes on housing price changes. According to hedonic price theory, the housing market may be used to estimate the implicit prices of clean air and the economic value of pollution reductions to individuals. A statistically significant negative relationship between changes in property values and

pollution levels across counties is interpreted as evidence that clean air has economic benefits. (For a summary of the theory, you could read: Rosen, Sherwin, "The Theory of Equalizing Differences," Chapter 12 in Handbook of Labor Economics, Volume 1, 1986, pp. 641-92.)
A basic model for the change in housing prices at the county level could be: Change in housing price = g(economic shocks, changes in county characteristics, change in air pollution)

1. Estimate the relationship between changes in air pollution and housing prices, both not adjusting and adjusting for other housing price determinants (use pop7080 as weights). What do your estimates imply and do they make sense? Describe the potential omitted variables biases. What is the likely relationship between economic shocks and pollution and housing price changes? Using the observable measures of economic shocks (dincome, dunemp, dmnfcg), provide evidence on this.

2. Suppose that federal EPA pollution regulation is a potential instrumental variable for pollution changes during the 1970s. What are the assumptions required for mid-decade regulatory status (tsp7576) to be a valid instrument for pollution changes when the outcome of interest is housing price changes? Provide evidence on the relationship between the regulatory status indicator and the observable economic shock measures. Interpret your findings.

3. Document the first-stage relationship between regulation and air pollution changes and the reduced form relationship between regulation and housing price changes, both not adjusting and adjusting for other covariates. Interpret your findings. How does two-stage least squares use these two equations? Now estimate the effect of air quality changes on housing price changes using two-stage least squares and the tsp7576 indicator as an instrument (not conditioning and conditioning on other observables). Interpret the results.

4. In principle, the regulation indicator variable should be a discrete function of pollution levels in 1974. Specifically, the EPA is supposed to regulate those counties in 1975 who had either an annual geometric mean of TSPs above 75 units (mg/m3) or a 2nd highest daily concentration above 260 units in 1974. Based on this notion, redo part c) using mtspgm74 as an instrument for pollution changes. Interpret your findings.

5. Describe how one could use this discontinuity in treatment assignment to derive an alternative estimate of the capitalization of pollution changes. Under what conditions will this estimate be valid? Based on this concept, estimate the nonparametric bivariate relationships between pollution changes and 1974 TSPs levels and housing price changes and 1974 TSPs levels. Do this using the lowess STATA command or a similar command in R (experiment with relatively small bandwidths between 2-4). Plot the two conditional mean functions on either side of the discontinutiy (changes in air quality for discrete 1974 values on either side of threshold). Interpret your findings and relate them to the results in (4).

6. Now use linear regression to estimate the predicted change in housing prices based on the control variables (excluding the change in TSPs). This provides a single-index measure of the housing price changes predicted to occur due to other variables changing. Estimate the nonparametric bivariate relation between this index and 1974 TSPs levels and plot it against the smoothed housing price changes from (5). Explain how this provides an indirect test of the smoothness condition required by the regression discontinuity design and interpret your findings.

7. A number of counties with annual geometric mean TSPs below 75 units in 1974 were regulated due to having as few as 2 bad days. This implies that we can compare regulated and

unregulated counties with identical average TSPs levels in the regulation selection year (below 75 units). For those counties with 1974 mean TSPs between 50 and 75 units, estimate the bivariate relation between TSPs changes and 1974 TSPs levels separately for regulated and unregulated counties and plot the results. Now do the same for the relation between log-housing price changes and 1974 TSPs levels. Interpret your findings. Since there are fewer observations, you may need to use bigger bandwidths than those in part (5) (e.g., bandwidths between 6-9).

8. Under what assumptions will two-stage least squares identify the average treatment effect (ATE)? If ATE is not identified, describe what may be identifiable with two-stage least squares estimation. Under what conditions is this effect identified? Give some intuition on what this effect may represent when one uses EPA regulation as an instrument variable.

9. Provide a concise synthesis/summary of your results. Discuss the credibility of the various research designs underlying the results.