

## Problem Set

This exercise examines the following research question: What is the relationship between changes in air pollution and housing prices?<sup>1</sup> For background on the topic and the data source refer to [Chay and Greenstone \(2005\)](#): “Does Air Quality Matter? Evidence from the Housing Market,” *Journal of Political Economy*, April 2005, 376-424. We will analyze the poll7080.dta data set.

- **Data Source:** This STATA data extract (poll7080.dta) is 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.

- **Data Notes:**

1. There are 1,000 observations 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).  
tsp75 = indicator equal to one if the county was regulated by the Environmental Protection Agency (EPA) in 1975 and equal to zero, otherwise.  
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.  
mtspgm75 = annual geometric mean of TSPs in 1975.
3. The other relevant variables are:  
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, vacant70 and vacant80 = housing vacancy rate in 1970 and 1980, vacrnt70 = rental vacancy rate in 1970, 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, deduc = change in fraction of spending on education, dhghwy = change in % spending on highways, dwelfr = change in % spending on public welfare, dhlth = change in % spending on health, blt1080 = % of houses built in the last 10 years as of 1980, blt2080 = % of houses built in the last 20 years as of 1980, bltold80 = % of houses built more than 20 years ago as of 1980.

The remaining variables in the data set are polynomials and interactions of the control variables.

---

<sup>1</sup>Adapted from Ken Chay’s Econ 2320 course materials.

- **Research Question:** 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 should read [Rosen \(1984\)](#) sections 1-3. A basic model for the change in housing prices at the county level could be:

Change in housing prices =  $g(\text{economic shocks, changes in county characteristics, change in air pollution})$

Please complete the following questions and provide a concise summary of your empirical results when appropriate. I recommend that you use Stata to complete this assignment, but R, python, or Matlab are also acceptable. You should submit your writeup as a pdf as well as your code in a .zip file on Canvas. Wherever possible, you should adhere to the coding guidelines in [Code and data for the social sciences: A practitioner's guide](#). Typeset your results with Latex or markdown software. Your written responses may be in bullet points and should include all relevant tables and figures.

1. Estimate the relationship between changes in air pollution and housing prices: 1) not adjusting for any control variables; 2) adjusting for the main effects of the control variables listed on the previous page; and 3) adjusting for the main effects, polynomials and interactions of the control variables included in the data set. 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, dden, durban, blt1080), 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 1975-1976 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 (tsp7576) and air pollution changes and the “reduced-form” relationship between regulation and housing price changes, using the same three specifications you used in part a). 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 for the three specifications. Interpret the results. Now do the same using the 1975 regulation indicator, tsp75, as an instrumental variable. Compare the findings.
4. In principle, the 1975 regulation indicator variable, tsp75, 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 ( $\mu\text{gm}^{-3}$ ) or a 2nd highest daily concentration above 260 units in 1974. Use this discontinuity in treatment assignment to derive alternative estimates of the capitalization of pollution changes. Under what conditions will these estimates be valid? Use a graphical analysis to examine the validity of these conditions.

5. Based on your reading of the Rosen Handbook chapter, describe the theoretical reasons why the effects of pollution changes on housing price changes may be heterogeneous. Write down a simple statistical model that allows for this possibility. Under what assumptions will two-stage least squares identify the average treatment effect (ATE)? What is the economic interpretation of ATE in the context of hedonic theory? If ATE is not identified, describe what may be identifiable with two-stage least squares estimation. Give some intuition on what this effect may represent when one uses EPA regulation as an instrument.
6. Now use a “Garen-type” control function approach to estimate the effect of pollution changes and test for heterogeneity in the effects using `tsp7576` as the instrumental variable. Do this for all three specifications and use the bootstrap (`bstrap` or `bsample` command in Stata) to calculate the correct standard errors. Interpret your findings in the context of hedonic theory. Under what conditions does this approach identify the average treatment effect?
7. Concisely and coherently summarize all of your findings. In this summary, describe the estimated effects of air pollution on housing price changes, discuss the “credibility” of the research designs underlying the results, and defend whether you think your “best” estimate of the effect of pollution is credibly identified. State why or why not.