"A Rational Expectations Approach to Hedonic Price Regressions with Time-Varying Unobserved Product Attributes: The Price of Pollution"

By

Patrick Bajari
Jane Cooley Fruehwirth
Kyoo il Kim
Christopher Timmins

Data Documentation and Estimation Code

This document describes how to go about obtaining the data we use in our estimation. It also describes the accompanying Matlab and Stata files which can be used to estimate the models in the paper.

(1) Housing Transactions Data from DataQuick

Our housing transactions data were purchased from DataQuick Information Systems (www.dataquick.com) as part of an agreement with the Department of Economics at Duke University. That agreement prohibits us from releasing any of the housing data used in this study. The data can, however, be purchased from DataQuick.

To facilitate replication of our data set, DataQuick has permitted us to upload the property identification numbers for the houses that we use in our estimation. We are not allowed to disclose any additional information about these houses. The attached data files contain property id's for our main estimation sample (data2sales.txt) and for our test of the efficient markets hypothesis (data3sales.txt). Each row of data2sales.txt contains the following variables:

id pmc2 soc2 ozc2 pmc1 soc1 ozc1

"id" refers to the property id assigned by DataQuick. "pmc" refers to particulate matter concentration, "soc" refers to sulfur dioxide concentration, and "ozc" refers to ozone concentration. "1" and "2" refer to the first and second sale associated with each house. data3sales.txt simply adds information for the third sale.

In order to replicate our full data set, one must contact DataQuick and request the sample of residential (i.e., non-commercial) housing for the counties and years described in the text of the paper. One will need to find the two sales that correspond to each property id in data2sales.txt and the three sales that correspond to each property id in data3sales.txt. From the proper record, one can find each of the data fields we use in estimation, including the year in which the sale took place.

Note that when data are purchased from DataQuick, they describe a particular set of counties over a particular time period. Importantly, all transactions recorded in each county up until the time of purchase are included in the file. As such, Bay Area data purchased in the future will contain more transactions than the data we used in this paper. This raises another important feature of DataQuick data. In particular, when DataQuick records the attributes of a house, it records the attributes measured at the time of the most recent transaction. This means that, as houses repeatedly transact over time, their attributes may change if they are upgraded or depreciate in a significant way. As such, it may be difficult match exactly a DataQuick data set purchased in the future with that used in our analysis. We would expect, however, the method to be equally applicable to a new data set.

(2) California Criteria Air Pollution Data

For reference purposes, the pollution data were obtained from the California Air Resources Board's website:

http://www.arb.ca.gov/aqd/aqdcd/aqdcddld.htm

At this website, one can also find file documentation along with definitions for each pollution measure that is available. In particular, see the files LOCATION.XLS and YSMULTIC.XLS. We employed the data contained on California Air Resources Board's DVD February 2008.

We provide the pollution levels associated with each property id in data2sales.txt and data3sales.txt.

(3) Estimation Code

See the attached Matlab and Stata files described below (9 total files included), which run the regressions and calculate the standard errors in our paper. These estimation algorithms are applicable to any housing transactions and amenities data organized in a similar fashion to our DataQuick data.

A. Cross-Section Model (Stata "do" file)

A1. crosssection.do (run estimations for the cross-section models)

B. House Fixed Effect Model (Stata "do" file)

B1. genyeardummy.do (generate year dummies)

B2. genfixedeffects.do (generate interaction variables with fixed effects)

B3. runfixedeffects.do (run estimations for the fixed effect models)

C. Efficient Housing Market Model (MATLAB "m" files)

C1. bckt_hedonic_2snls.m (main program to run estimations and report standard errors)

C2. secondstepall_2sls_fullgamma_obj.m (sub program to call objective function for minimization (two stage nonlinear LS) with all pollutants)

C3. secondstepa_2sls_fullgamma_obj.m (sub program to call objective function for minimization with PM10 only)

C4. secondstepb_2sls_fullgamma_obj.m (sub program to call objective function for minimization with SO2 only)

C5. secondstepd_2sls_fullgamma_obj.m (sub program to call objective function for minimization with O3 only)

Note: additional details and instructions are provided inside each program