

Econ 26610 Problem Set 2

Prepared by Feng Lin*

Due 2025-02-24

1 Hedonic Regression

As Professor Rossi-Hansberg discussed in class, “Neighborhoods in Bloom” (NiB) is a program in Richmond, VA between 1999 and 2004 that subsidizes acquisition, demolition, rehabilitation, and new construction in selected disadvantaged neighborhoods. We obtained a geo-coded listing of all properties sold between 1993 and 2004 in the area with various information on their characteristics and a list of “funding centers” (convex combinations of subsidized sites within each affected neighborhood). We are interested in examining the housing externalities due to the NiB program.

Please download data files for this question from Canvas Modules.

- `data_pre.csv` and `data_post.csv`: Listings of selected properties sold in the pre-policy period 1993–1998 and the post-policy period 1999–2004.
 - `data_fundingcenters.csv`: The location of funding centers.
 - Please refer to Table 1 for a list of variables in the data files and Table 2 for a list of neighborhoods.
1. The data set only includes properties that did not receive NiB funding. Please explain why we may want to use this subsample instead of the full sample for the hedonic valuation.
 2. In the combined sample, what is the share of properties with each type of heating? What are the average age and the average size of land area?
 3. We would like to estimate the log price of land per square foot from a hedonic regression of the following form

$$p_i = \beta_0 + \beta_1 Z_{1i} + \cdots + \beta_K Z_{Ki} + \varepsilon_i,$$

*Please send questions/corrections to fenglin2@uchicago.edu.

where p_i is the log sale price per square foot of land for property i and $Z_{1i}, \dots, Z_{Ki}$ are characteristics of property i that include indicator variables for sale year (base = 1998 in the pre-policy period and 2004 in the post-policy period), indicator variable for central AC, indicator variables for exterior type (base = wood and other), indicator variables for heating type (base = gas heating), square footage of living area, age, square footage of land area, indicator variables for property condition (base = good), and numbers of bathrooms. Please implement this regression and present the regression results for the pre- and post-policy period separately.

4. Please answer the following questions regarding the regression:
 - (a) For the indicator variables, why do we need to choose a base category and exclude it from the regressors?
 - (b) Why do we want to run regressions separately for the pre- and post-policy periods?
 - (c) Please interpret the coefficients on the indicator variable for having central heating.
 - (d) Please interpret the coefficients on the square footage of living area.
 - (e) Are coefficients on the square footage of land area positive or negative? What is a possible explanation? *[Hint: Think about the left-hand-side variable.]*
 - (f) After running the regressions, how do we obtain the estimates for log price of land per square foot?
5. What is the average distance of the selected properties to the closest funding center in the corresponding neighborhood for the whole sample? Please create a histogram of these distances for each neighborhood, using a bin size of 50 feet.
6. Plot a local polynomial smooth of log land price per square foot on the distance to the closest funding center in the corresponding neighborhood. Please use the following specifications:
 - Make one plot or panel for each affected neighborhood.
 - Overlay the smoothed curves for the pre-policy and post-policy periods.
 - Set the limit of x -axis to $[0, 1500]$.
 - Take a look at `ggplot2::geom_smooth` of R or `twoway lpoly` in Stata for commands to make the smoothed curves. (For `ggplot2::geom_smooth`, you need to use “loess”.)
7. Based on the figure in the previous question, what is the approximate increase in land value for properties right next to the funding centers in Highland Park after NiB in comparison to before? What is the approximate increase for properties around 1,500 feet away from the funding centers in Highland Park? Are similar patterns present in other neighborhoods? What can you say about the effects of NiB based on these patterns? *[Hint: You can read off the approximate values visually.]*

8. *[BONUS QUESTION]* In the paper that studies housing externalities due to NiB, the authors use a non-parametric regression to estimate “smoothed” log land price per square foot at grid points corresponding to properties in the pre-policy sample. This method essentially calculates log land price per square foot at each grid point as a weighted average of the value at adjacent properties. By doing this, the authors are able to obtain figures on page 30 of Week 3 Slides that show very sharp gradient of price increase by distance to the funding centers. Compare the authors’ procedure to what we do in Q1.5 and discuss why they may be able to produce sharper results.

Table 1: List of Variables

Variable	Description
geo_x	Cartesian coordinate x
geo_y	Cartesian coordinate y
neighborhood	Neighborhood
lpricesqft	$\log(\text{Real sale price per square foot of land})$
lprice	$\log(\text{Real sale price})$
year	Year of sale
landarea	Land area in square foot
livingarea	Living area in square foot
age	Age
condition	Condition
exterior	Exterior type (brick, vinyl, or other)
centralac	Whether the property has central AC
heating	Heating type (central, gas, or hot water)
bath	Number of bathrooms

Table 2: List of Neighborhoods

Code	Description
CH	Church Hill
BL	Blackwell
HP	Highland Park and South Barton Heights
JW	Jackson Ward and Corver
BM	Bellemeade (Control)

2 Look for Spatial Sorting of Workers

We are interested in examining if there is spatial sorting of workers according to particular characteristics. We will consider a series of empirical exercises and think about what they tell us about the sorting of workers.

1. When we talk about the spatial sorting of workers, what phenomenon do we have in mind? Please use the spatial sorting of workers based on education as an example.
2. Using worker-level data from Spain, we estimate the following relationship:

$$w_{ict} = \sigma_c + \beta_1 x_{1it} + \cdots + \beta_K x_{Kit} + \eta_{ict},$$

where

- w_{ict} denotes worker i 's log wage in city c at time t ;
- σ_c denotes city fixed effect for city c ;
- $x_{1it}, \dots, x_{Kit}$ are observable controls that include worker's experience, worker's firm tenure, indicator variables for skill groups of worker's occupation, indicator variables for worker's education level;
- η_{ict} denotes the error term.

What does σ_c capture in this specification? If $\sigma_1 = 2\sigma_2$, how much higher would the average wage in city 1 be compared to city 2, conditional on the observable controls?

3. Regressing σ_c that we obtained from the previous step on log city population, we get an coefficient of 0.0455. How much higher would the average wage be if city population triples?
4. Can the sorting of workers on unobserved ability lead to this correlation? Are there any other possible explanation for this correlation? Can we tell these channels apart with the given regression?
5. Consider an alternative specification:

$$w_{ict} = \sigma_c + \mu_i + \beta_1 x_{1it} + \cdots + \beta_K x_{Kit} + \eta_{ict},$$

where

- μ_i denotes worker fixed effect for worker i ;
- The other variables can still be describes as above.

What does μ_i capture in this specification? Please provide some examples.

6. Regressing σ_c that we obtained from the alternative specification on log city population, we get an coefficient of 0.0241. Compare this with the baseline results. Does the difference suggest that spatial sorting of workers based on unobserved characteristics is important? What are some potential confounding factors related to learning?
7. Suppose that there is no spatial sorting of worker based on unobserved ability, what would you expect to be the difference between the distribution of μ_i of workers in a large city and that in a small city? *[Hint: It may be helpful to assume that μ_i follows a bell curve in the small city and discuss how the distribution in the large city would be different from this baseline. A figure can be useful.]*

3 Road Congestion

Consider a simple model of commuting that helps us understand the effect of building more roads on Vehicle-Kilometers Traveled (VKT). **Please make sure that you answer each sub-question and present your answer clearly.**

1. What does Vehicle-Kilometers Traveled measure? Is it a measure at the aggregate level or individual level? If there are 500,000 vehicles in a city and each of them travels 5,475 kilometers annually, what is the annual VKT for the city?
2. Is VKT a direct measure of congestion? What other information (if any) do you need to make inference about congestion in a city using VKT?
3. Let $P(\text{VKT})$ be the inverse demand for VKT in a city. What does $P(\text{VKT})$ capture conceptually? Is $P(\text{VKT})$ increasing or decreasing in VKT?
4. Let R denote lane kilometers of roads in the city. Let $AC(\text{VKT}|R)$ be the average cost of VKT given roads R . What does $AC(\text{VKT}|R)$ capture conceptually? Why would you expect $AC(\text{VKT}|R)$ to be increasing in VKT?
5. What is the equilibrium level of VKT*? A figure may be useful.
6. If floods caused by a hurricane permanently destroyed some roads, how would R change? How would $AC(\text{VKT}|R)$ shift as a result? Why? How would the equilibrium level of VKT* change accordingly? What is the interpretation?
7. As you have seen in class, research has shown that 1% of new highway lane kilometers is associated with 1% more VKT. What does this say about congestion? Assume that $AC(\text{VKT}|R) = AC(\text{VKT}/R)$, what does this say about the slope of the $P(\text{VKT})$ curve? Please be strict with your reasoning. Please use a figure to make your argument strict.

4 Short Answers

Answer the following questions with less than 5 bullet points each:

1. What issue does the dartboard approach help address when measuring industry concentration?
2. Why should we care about input-output linkages when understanding the welfare effects of local shocks?