House Price “Hedonic” Regressions

\[ \ln(price_{iag}) = \alpha + x_{iag} \beta + z_j \delta + test_{aj} \gamma + \varepsilon_{iag} \]

- \( i = house, a = school, j = district \)
- \( x_{iag} = house\) characteristics
- \( z_j = neighborhood(district)\) characteristics
- \( test_{aj} = test\) score for local \( k - 6\) school

Omitted variable bias

- Incomplete measurement of neighborhood characteristics
- These may be correlated with both price and test score
  - Example: high income people live in high-scoring schools areas and neighborhoods with nice parks, sidewalks, quiet, etc.
- Test score could be capturing these other positive characteristics – biasing the coefficient up
Solution

- Exploit the fact that on school boundaries homes on either side of the street are in the same neighborhood but in different schools
- How does the solution impact internal and external validity?

New model

\[ \ln(price_{iab}) = \alpha + x_{iab}\beta + test_{ai}\gamma + \lambda_{ib} + \epsilon_{iab} \]

- \(i\) = house, \(a\) = school, \(b\) = boundary
- \(x_{iab}\) = house characteristics
- \(test_{ai}\) = test score for local \(k - 6\) school
- \(\lambda_{ib}\) = boundary fixed effect

New model

\[ \ln(price_{iab}) = \alpha + x_{iab}\beta + test_{ai}\gamma + \sum_{b=1}^{B-1} D_{b}\theta_{b} + \epsilon_{iab} \]

- \(B\) boundaries in total
- \(D_{b} = 1\) if from boundary \(b\)
- \(= 0\) otherwise
Important questions

• What does the dummy variable $D_h$ capture?

• Why do we include B-1 dummies?

Data

• All home sales for Middlesex, Essex and Norfolk counties, 1993-95
  – Why is MA a good place for this study?

• Single-family residences

• Districts with at least two schools

• Eliminate districts with school choice. Why?

• MA Educ. Assessment Program (MAEP)
  – Given in grades 4, 8, 12 in 5 subjects
  – Author uses sum of reading + math

Final Sample

• 22,679 single-family home sales

• 39 districts

• 181 attendance boundaries

TABLE 1

Summary Statistics

<table>
<thead>
<tr>
<th>Distance from boundary:</th>
<th>Full sample</th>
<th>0.35 mile</th>
<th>0.50 mile</th>
<th>0.75 mile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>Mean</td>
<td>S.D.</td>
</tr>
<tr>
<td>House price ($) (1993)</td>
<td>180,076</td>
<td>115,928</td>
<td>185,769</td>
<td>190,064</td>
</tr>
<tr>
<td>Adjusted*</td>
<td>12.1</td>
<td>0.5</td>
<td>12.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>3.2</td>
<td>0.9</td>
<td>3.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>1.5</td>
<td>0.7</td>
<td>1.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Age of building</td>
<td>53</td>
<td>30</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Less than 10000sq.</td>
<td>17.3</td>
<td>15.0</td>
<td>14.3</td>
<td>12.5</td>
</tr>
<tr>
<td>Internal square footage</td>
<td>1.8</td>
<td>0.8</td>
<td>1.8</td>
<td>0.6</td>
</tr>
<tr>
<td>School characteristics*</td>
<td>27.6</td>
<td>1.4</td>
<td>27.5</td>
<td>1.4</td>
</tr>
</tbody>
</table>
Some questions to ponder?

- Why do the standard errors increase as we move from models (1) to (2)?
  - There are two reasons

- Suppose you wanted to test whether the coefficients on test are different in models (1) and (2) – how would you do it?