Using Administrative Data to Understand Labor Markets

David Card
April 2022

All results from LEHD in this presentation have been reviewed for disclosure avoidance by the Census Bureau (DRB clearance numbers CBDRB-FY21-122, CBDRB-FY21-149, CBDRB-FY21-288). All opinions are our own.
A little history

• up until the 1960s – government agencies and researchers mainly relied on censuses and cross sectional surveys

• early surveys (Current Population Survey, CPS, 1939+) revolutionized our understanding of “unemployment”

• analysis of Census data (Miller, 1955; 1966) revealed the distribution of income, relationship of income and education

• public versions of CPS and Census – basic inputs for modern labor economics
A little history (2)

- mid-1960s: idea of constructing a “Panel Survey” – return to the same households/people in successive years
- Panel Study of Income Dynamics (PSID), National Longitudinal Surveys (NLS)

- longitudinal surveys revealed many new “facts”
  - instability in families
  - permanent and transitory components of earnings
  - impacts of job mobility, long tenure jobs...
A little history (3)

By the early 1990s, limitations of standard and panel surveys were becoming clearer

1. response problems (undercount/nonresponse/attrition)

2. lack of granularity ('upper tail of income', narrow geographies, rare populations)

3. “network” features – parent/child; worker/co-worker
the ‘administrative data’ revolution

- tax records are created automatically for purposes of recording income flows (W2), monitoring program eligibility (SSA), and keeping track of program charges (Workers Comp, UI)

- “frame” of records is sometimes clear; but can be messy (Abowd, McKinney, Zhao, 2017 – presence of ‘bad’ SSN’s in LEHD)

- create network links
  - worker/firm, worker-coworker
  - parent/child (tax returns)
the ‘administrative data’ revolution (2)

“Network” effects

- key advantage of a “universe” admin record data base is the ability to see the links between units
- prominent example: work by Chetty et al documenting intergenerational correlation in incomes of parents and children at the county level
- for labor economists, the #1 link of interest is “worker-firm”
- for urban economists, #1 link of interest is “worker-place”
network interactions

“Network” effects (aside on matching models)
- older panel surveys (PSID, NLS) – we could see that changing jobs mattered, but could not identify the specific employer
- led to development of models of *idiosyncratic matching* – each job has a “match effect” (like a marriage model)
- you could go to Berkeley or to Stanford and get a new match effect. **No common component for everyone at Stanford**
- still a dominant perspective – analytically convenient, and aligns with the idea of “comparative advantage”
network interactions (2)

“Network” effects (worker-firm)
- when we can see all (or nearly all) the job market links: we can separate out the contributions of
  - worker side
  - employer side
  - possible match effects
Allows us to ask:
  - does moving from Berkeley to Stanford raise your wage?
  - or would Stanford people earn more at all jobs?
  - or do only some benefit from Stanford, others from Berk.
network interactions (3)

Spatial “Network” effects
- if we see workers as they move from city to city, we can separate out the contributions of
  - worker side
  - city effect
  - possible match effect
Allows us to ask:
  - does moving from St Louis to SF raise your wage?
  - or do people in SF have higher earnings in all cities
  - or do some people benefit from SF, others from St Louis
network interactions (4)

Abowd Kramarz Margolis (a.k.a. “mover design”)

\[
\text{wage (person } i, \text{ period } t) = \\
\text{person effect (fixed)} \\
+ \\
\text{employer effect (“add factor”)} \text{ << or city effect >>} \\
+ \\
\text{noise}
\]
network interactions (5)

“Network” effects (Mover design)
- when worker moves from firm 1 (add factor $\psi_1$) to firm 2 (add factor $\psi_2$) expect an average change of $\psi_2 - \psi_1$
  move from 2 to 1: $\rightarrow \psi_1 - \psi_2$
- if Berkeley $\rightarrow$ Stanford = +20%
  then Stanford $\rightarrow$ Berkeley = −20%
- common add factor $\leftrightarrow$ no “comparative advantage”
network interactions (6)

“Network” effects (Mover design)
Essentially a 2 step model:
Step 1: using movers, estimate the add factors for employers
(or cities)
Step 2: given the add factors, choose a worker effect that fits the mean of the “add-factor adjusted” wages from every period. This incorporates stuff we could in principle measure (education, gender) + stuff that matters but is hard to measure (motivation, ...
network interactions (7)

“Network” effects
What have we learned in past 20 years?
1. simple “add factor” model is surprisingly good
2. worker permanent component is large
3. firm add factors contribute 10-20% of variance
4. workers with larger permanent component tend to work at firms with larger add factor $\Rightarrow$ inequality magnified
5. disadvantaged workers (nonwhite, female) tend to work at firms with smaller add factors $\Rightarrow$ “gaps” magnified
Part 2: using LEHD to study CZ (Commuting Zone) effects

- with Moises Yi (Census Bureau) and Jesse Rothstein (Berkeley, California Policy Lab)
- use LEHD 2010 – 2018 + “mover design” (AKM)
- estimate slightly more complicated model:
  \[ \text{wage} \left( i, t \right) = \text{person effect (fixed)} + \text{add factor for (CZ \times industry)} + \text{age/time + noise} \]

Why CZ × industry?  
- industry matters  
- potential ‘agglomeration effects’
Research questions

1. Are there causal place effects?

2. Are city effects really industry effects?

3. Are place effects similar for different industries (i.e., are there industry × place match effects)?

4. Why are wages higher in larger cities?

5. Do place effects differ for low/high educated groups?

6. Do housing costs offset place-related earnings premia?
Research questions & summary of answers

1. Are there causal place effects?
   Yes, but most wage variation across CZ’s is “worker quality.”

2. Are city effects really industry effects?
   No. CZ and industry are almost separate add factors!

3. Are industry effects the same in different CZ’s?
   Nearly the same. Slightly larger industry effect when the industry is bigger.

4. Why are wages higher in larger cities?
   2/3 unobserved worker skill; 1/3 causal effects.

5. Do place effects differ for low/high ed groups?
   No – but higher-ed workers are more likely to congregate in certain CZs.

6. Do housing costs offset place-related earnings premia?
   Real earnings are lower in high-earning & large cities.
<table>
<thead>
<tr>
<th></th>
<th>All LEHD Observations</th>
<th>Estimation Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Share Quarterly Earnings $\geq$ $3800$</td>
<td>0.837</td>
<td>1</td>
</tr>
<tr>
<td>Mean Earnings (if $\geq$ $3800$)</td>
<td>16,050</td>
<td>16,600</td>
</tr>
<tr>
<td><strong>Number of CZ's during Sample Period:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 CZ</td>
<td>0.63</td>
<td>0.73</td>
</tr>
<tr>
<td>2 CZ's</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td>3+ CZ's</td>
<td>0.14</td>
<td>0.07</td>
</tr>
<tr>
<td><strong>Number of 2-digit Industries during Sample Period:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Industry</td>
<td>0.49</td>
<td>0.63</td>
</tr>
<tr>
<td>2 Industries</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>3+ Industries</td>
<td>0.24</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Motivation: Earnings before and after moves

• Divide CZs into quartiles (unweighted) by mean earnings.

• Event study sample

• Earnings adjusted for age, calendar time.
Motivation: Earnings before and after moves

- Divide CZs into quartiles (unweighted) by mean earnings.
- Event study sample
- Earnings adjusted for age, calendar time.
Post-estimation results

1. At the person-quarter level: CZ-industry effects add only small share of variance (but higher-wage workers are sorted to higher CZ-industries, \( \rho = 0.2 \))

2. At the CZ level, about 50% of the variance in mean wages is due to the presence of higher-earning people and 20% to the combination of CZ-industry add factors. These two components are very strongly correlated \( \rho = 0.6 \) \( \rightarrow \) inequality effect

3. CZ’s with high average add factor are of 2 types:
   - bigger, high wage cities (NYC, SF, etc)
   - resource intensive places
Quartile of origin and destination

CZ Avg. Effect

- 4 to 4
- 4 to 3
- 4 to 2
- 4 to 1
- 1 to 4
- 1 to 3
- 1 to 2
- 1 to 1

Mean log wage (adjusted for age and year)

Post move
Distinguishing industry and place effects

Our model has (CZ × industry) add factors $\psi_{cj}$

1. When we look at these, we see

$$\psi_{cj} = \text{CZ effect} + \text{industry effect} + \text{‘match’ effect}$$

   (41%)   (46%)   (13%)

2. The add factor for industry $j$ in CZ $c$ is bigger if that industry has a larger share of all local employees (“agglomeration”)

3. But CZ’s have surprisingly similar composition – small agglom.

4. The (weighted average) add factor for CZ $c$ is not explained by industry composition or differences in local return to industries
Are there different place effects for high and low education workers?

- important finding in recent literature:
  \[ \text{EdGap} = \text{Mean Wage (college)} - \text{Mean Wage (HS)} \]
  \[ \text{EdGap}_c = 0.64 \times \text{mean log wage in CZ} \]

- many people in LEHD were interviewed in 2000 Census or ACS → get their education
- use the “with education” subsample to re-estimate separate models for college (13+ yrs) or HS (<=12 yrs) education
- the two sets of (CZ × industry) effects are very similar
Are there different place effects for high and low education workers? (2)

mean wage = mean of add-factors + “person effect”

But if the add-factors for college and HS are the same, that means differences in EdGap have to be due to widening or narrowing of the difference in mean “person effects” for college vs HS workers

Key finding: what appears as a very high “return to college” in SF is really due to the fact that the college workers in SF are very high wage college workers, while the HS workers are about average (“differential sorting by ability”)
Housing costs and real wages

How much of earnings premium in higher-wage / larger CZs is eaten up by housing costs?
**Simple summary: move to 100 log point bigger city (1.1 M to 3 M people)**

<table>
<thead>
<tr>
<th></th>
<th>All workers</th>
<th>College Educ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean effect on earnings</td>
<td>+7.65%</td>
<td>+10.20%</td>
</tr>
<tr>
<td>Due to “person effects” of workers</td>
<td>+ 5.05%</td>
<td>+7.40%</td>
</tr>
<tr>
<td>Due to causal effect of CZ</td>
<td>+2.60%</td>
<td>+2.80%</td>
</tr>
<tr>
<td>Mean effect on housing costs</td>
<td>+ 18%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(based on rent info in ACS)</td>
<td></td>
</tr>
<tr>
<td>Mean effect on real wage</td>
<td>+ 2.60 – 18/3</td>
<td>+ 2.80 – 18/3</td>
</tr>
<tr>
<td>(assuming 1/3 housing share)</td>
<td>= – 3.4%</td>
<td>= – 3.2%</td>
</tr>
</tbody>
</table>
Conclusions

- Administrative data provide the “missing links” allowing us to connect children to parents, workers to co-workers, workers to firms...

- the basic “working model” (AKM):
  \[ \text{wage} = \text{worker’s permanent effect} + \text{“add factors”} \]

- widely applied to worker/firm
- also works for worker/place
- many new insights