Migration stats via ceteris paribus statistics

Ben Klemens

Office of Tax Analysis
U.S. Treasury

ASSA, 3-5 January 2020

https://tinyurl.com/subsetstable

Preliminary. Any opinions and conclusions expressed herein are those of the author and do not necessarily represent the views of the Department of the Treasury.
Motivation

• One in 20 households move over 80km: a major life change.
  ▶ More common than tax literature favorites like retirement or tuition
• Data (2 mins)
• Methods (5 mins)
  ▶ Test hypotheses via pseudo-controlled experiments
  ▶ We want a method with $\sim$zero degrees of freedom in design
    ▶ Model design for kitchen-sink regressions is delicate
• Results (8 mins)
  ▶ Open the firehose of stats about movers and different characteristics
The data
• Compiled by Chetty, Friedman, Yagan, and IRS counterparts.
• The U.S. formal economy, 2001–2015: 1,748,802,270 household observations, 82,711,474 moves.
Covariates for our tabulation

- marital status
- # of kids
- Schooling status (presence of 1098-T)
- Retirement
- Mortgage
- Income
- Employed/unemployed
- Local taxes
- Age
- Sex if unmarried
- Year

Also: population density, local unemployment, and housing costs.
Methods
Risk ratio (relative risk) review

- What is $P(\text{mv}|\text{mar})/P(\text{mv}|\text{single})$?

<table>
<thead>
<tr>
<th></th>
<th>Married</th>
<th>Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moved</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Stayed</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

$RR \equiv$ move risk given married vs move risk given not:

\[
\frac{A}{A+C} = \frac{A(B + D)}{B(A + C)}
\]
Run pseudo-experiments like it’s 1956

- Find only people with a mortgage and kids. What is $RR(mv|mar \text{ vs single})$?
- Find only people with a mortgage and no kids. What is $RR(mv|mar \text{ vs single})$?
- ...
- Report an aggregate somewhere in between
Subset instability

A made up example:

- For all: *ceteris paribus* \( RR(mv|\text{mar vs single}) = 118.4\% \)
- For subset with mortgage: *c.p.* \( RR = 110\% \)
- For subset without mortgage: *c.p.* \( RR = 115\% \)
Subset instability

A made up example:

• For all: ceteris paribus $RR(\text{mv}|\text{mar vs single})=118.4\%$
• For subset with mortgage: c.p. $RR=110\%$
• For subset without mortgage: c.p. $RR=115\%$
Let’s consider alternatives

They should be subset stable.
Let’s consider alternatives

They should be subset stable.

**Admissible**

- Smooth: all derivatives exist.
- Statistic can’t be zero or $\infty$ for every real-world data set.
  - If $B_{1000} = 0$, $\frac{1}{B_1} \cdot \frac{1}{B_2} \cdot \frac{1}{B_3} \cdot \frac{1}{B_{1000}} = \infty$
The Cochran-Mantel-Haenszel (CMH) statistic

Mantel and Haenszel (1956)

\[ A_i + B_i + C_i + D_i = 1, \text{ assign weight } w_i: \]

\[
\frac{\sum_i w_i A_i (B_i + D_i)}{\sum_i w_i B_i (A_i + C_i)}
\]
The CMH statistic is the *unique* admissible aggregate risk ratio satisfying subset stability.

CMH statistic ⇒

Anything else ⇒
A warm-up test
Percent of . . . who move

<table>
<thead>
<tr>
<th>Status</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, single</td>
<td>5.1%</td>
</tr>
<tr>
<td>Female, single</td>
<td>6.1%</td>
</tr>
<tr>
<td>Unmarried</td>
<td>5.5%</td>
</tr>
<tr>
<td>Married</td>
<td>3.7%</td>
</tr>
</tbody>
</table>
Claims

- $H_0$: Being married has no effect on a household’s $P(\text{moving})$.
- $H_1$: Being married lowers a household’s $P(\text{moving})$.
- Similarly for being a single man relative to being a single woman.
Claims

- $H_0$: Being married has no effect on a household’s $P(\text{moving})$.
- $H_1$: Being married lowers a household’s $P(\text{moving})$.
- Similarly for being a single man relative to being a single woman

But: Yes, both mortgage and kids reduce chance of moving.
- Married couples are more likely to have a mortgage and kids.
- 80.4% of singles with a mortgage are men.
**ceteris paribus** $P(\text{move}|\text{status})/P(\text{move}|\text{not status})$

<table>
<thead>
<tr>
<th>Status</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, single</td>
<td>107.0 %</td>
</tr>
<tr>
<td>Female, single</td>
<td>93.5 %</td>
</tr>
<tr>
<td>Unmarried</td>
<td>84.5 %</td>
</tr>
<tr>
<td>Married</td>
<td>118.4 %</td>
</tr>
</tbody>
</table>

Reject $H_0$ and $H_1$, for marrieds and single men.
Moving by AGI band

\( H_1 \): lower-income individuals move less than others
Moving by AGI band
Counterfactual income change
Simple counterfactual $\Delta$ AGI

- Separate into all Else Equal groups.
- For each group, what is movers’ AGI - stayers’ AGI?
- Stack the per-cell measurements.
Now just people leaving school
Now in table form

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>% of movers</th>
<th>$R + 2$</th>
<th>$R + 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>100 %</td>
<td>55.8 %</td>
<td>63.5 %</td>
</tr>
<tr>
<td>Leaving school, all</td>
<td>6.7 %</td>
<td>78.5 %</td>
<td>72.0 %</td>
</tr>
<tr>
<td>All others</td>
<td>93.3 %</td>
<td>53.0 %</td>
<td>61.9 %</td>
</tr>
</tbody>
</table>
[Exclude school leavers]

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>% of movers</th>
<th>$R + 2$</th>
<th>$R + 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retiring</td>
<td>0.7 %</td>
<td>32.5 %</td>
<td>34.8 %</td>
</tr>
<tr>
<td>Retired</td>
<td>1.5 %</td>
<td>40.2 %</td>
<td>45.0 %</td>
</tr>
</tbody>
</table>

%pos | median
---|---
−0.08 | −0.06
−0.03 | −0.02

Counterfactual income change
### General (younger) population

Not leaving school, not retiring, under 45, AGI < 100k

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>% of movers</th>
<th>$R + 2$</th>
<th>$R + 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%pos</td>
<td>median</td>
<td>%pos</td>
</tr>
<tr>
<td>Single men, no children</td>
<td>18.9%</td>
<td>62.9%</td>
<td>0.04</td>
</tr>
<tr>
<td>Single women, no children</td>
<td>13.3%</td>
<td>60.6%</td>
<td>0.03</td>
</tr>
<tr>
<td>Married, no children</td>
<td>5.5%</td>
<td>63.6%</td>
<td>0.04</td>
</tr>
<tr>
<td>Married, 1+ children</td>
<td>9.9%</td>
<td>61.7%</td>
<td>0.03</td>
</tr>
<tr>
<td>Single men, 1+ children</td>
<td>3.2%</td>
<td>51.6%</td>
<td>0.00</td>
</tr>
<tr>
<td>Single women, 1+ children</td>
<td>4.8%</td>
<td>44.9%</td>
<td>−0.01</td>
</tr>
</tbody>
</table>

**Klemens Counterfactual income change**
Summary slide

• Simple pseudo-experiments can be adapted to high-dimensional analyses
  ▶ They couldn’t do 13-dimensional crosstabs in the 1900s. We can now.
  ▶ No delicate model design debates. Article proves that even the aggregation has only one option.

• Hypothesis: people move to expand their income.
  ▶ Works great for people moving after part-time school or grad school.
  ▶ But we reject for almost half the population.

Me: ben.klemens@treasury.gov
The paper: https://tinyurl.com/subsetstable
CMH calculator (w/demo): https://github.com/b-k/cmh.py