Migration stats via ceteris paribus statistics

Ben Klemens

Office of Tax Analysis
U.S. Treasury

ASSA,3-5 Januay 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

Outline

- 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

The Data Bank

- 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 $P(mv|mar)/P(mv|single)$?

	Married	Single
Moved	A	B
Stayed	C	D

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

$$\frac{\frac{A}{A+C}}{\frac{B}{B+D}} = \frac{A(B+D)}{B(A+C)}$$

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* $RR(mv|mar \text{ vs } single)$=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* $RR(mv|mar \text{ vs } single)$=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$\equiv$

- Smooth: all derivatives exist.
- Statistic can't be zero or ∞ for every real-world data set.
  - If $B_{1000} = 0$, $\frac{1}{B_1} \cdot \frac{1}{B_2} \cdot \frac{1}{B_3} \cdot \frac{1}{B_1 000} = \infty$

The Cochran-Mantel-Haenszel (CMH) statistic

Mantel and Haenszel (1956)
$A_i + B_i + C_i + D_i = 1$, assign weight w_i:

$$\frac{\Sigma_i w_i A_i (B_i + D_i)}{\Sigma_i w_i B_i (A_i + C_i)}$$

Theorem (proven in this article)

The CMH statistic is the *unique* admissible aggregate risk ratio satisfying subset stability.

CMH statistic $\Rightarrow$ 

Anything else $\Rightarrow$ 

A warm-up test

Percent of ... who move

Male, single	5.1 %	Unmarried	5.5 %
Female, single	6.1 %	Married	3.7 %

Claims

- H_0: Being married has no effect on a household's P(moving).
- H_1: Being married lowers a household's P(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(moving).
- H_1: Being married lowers a household's P(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 $P(\textbf{move}|\textbf{status})/P(\textbf{move}|\textbf{not status})$

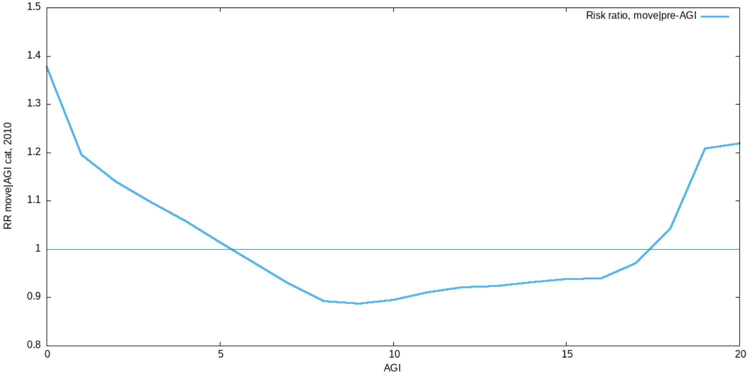
Male, single	107.0 %	Unmarried	84.5 %
Female, single	93.5 %	Married	118.4 %

Reject H_0 and H_1, for marrieds and single men.

Moving by AGI band

H_1: lower-income individuals move less than others

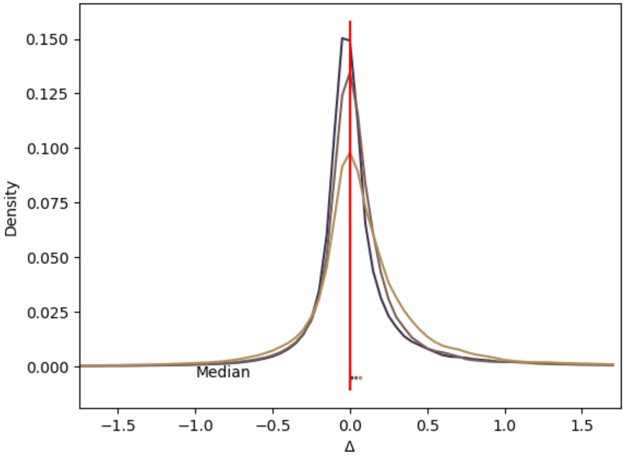
Moving by AGI band



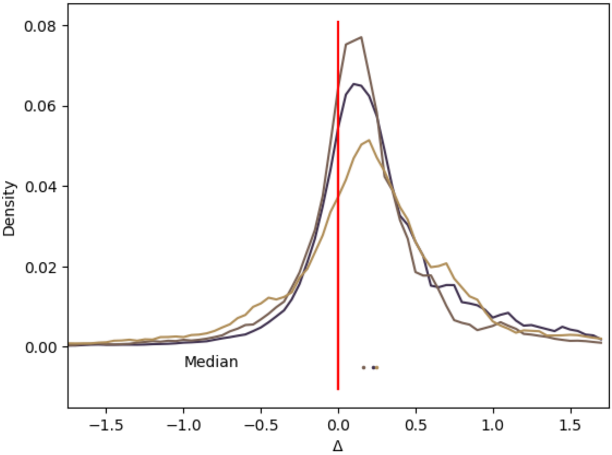
Counterfactual income change

Simple counterfactual Δ 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

Subgroup	% of movers	$R+2$ %pos	$R+2$ median	$R+10$ %pos	$R+10$ median
All	100 %	55.8 %	0.01	63.5 %	0.06
Leaving school, all	6.7 %	78.5 %	0.23	72.0 %	0.24
All others	93.3 %	53.0 %	0.00	61.9 %	0.05

Retirees

[Exclude school leavers]

		$R+2$		$R+10$	
Subgroup	% of movers	%pos	median	%pos	median
Retiring	0.7 %	32.5 %	−0.08	34.8 %	−0.06
Retired	1.5 %	40.2 %	−0.03	45.0 %	−0.02

General (younger) population

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

Subgroup	% of movers	$R+2$ %pos	$R+2$ median	$R+10$ %pos	$R+10$ median
Single men, no children	18.9 %	62.9 %	0.04	75.1 %	0.13
Single women, no children	13.3 %	60.6 %	0.03	71.7 %	0.13
Married, no children	5.5 %	63.6 %	0.04	74.0 %	0.15
Married, 1+ children	9.9 %	61.7 %	0.03	72.3 %	0.12
Single men, 1+ children	3.2 %	51.6 %	0.00	55.4 %	0.03
Single women, 1+ children	4.8 %	44.9 %	−0.01	56.7 %	0.03

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`