Weight Gains: Measuring Inequality in Panel Data and Other Adventures in Sampling

Andrew Foster

Brown University

July 20, 2019

Introduction

- Stratified sampling has played an important role in the collection of development data
- Among the most prominent data sets are stratified by income or landholding including ICRISAT and ARIS/REDS
- But sample stratification and weights are typically seen as something of a nuisance by empirical researchers
- But solutions are rarely mechanical–they interact importantly with substantive questions.
- Important for applied researchers to be more literate in sampling/weighting issues

Introduction II

Some recent intellectual adventures involving sampling and weights...in each case the focus was elsewhere but my coauthors and I had to confront these issues.

- Measuring changes in inequality from panel data
- Constructing estimates of economic mobility over multiple generations
- Evaluating changes in the spatial distribution in well-being as development proceeds
- Examining the effects of democratization.

Recombination

Constructing representative samples from panel data (Foster and Milusheva 2018).

- Data are typically collected from samples of households due to importance of the household in the allocation of resources
- Households are not fixed over time–they "recombine".
- Probability of observing particular individuals/household depends on endogenous behaviors inclusive of recombination
- Weights are thus endogenous with respect to important aspects of behavior. Is this something we need to account for in analyzing weighted data?

Matlab Data

- Vital registration data in a Demographic Surveillance System over 40 years with censuses in 1974 and 1982
- Economic panel survey in 1996, 2014.
- Panel survey sampled based on *bari* in 1996
- One objective was to measure effects of MCH-FP program starting in 1978.
- How do we handle the fact that the survey is representative in 1996 but antecedents are not representative in 1974 and descendants are not representative in 2014?

Representative sample

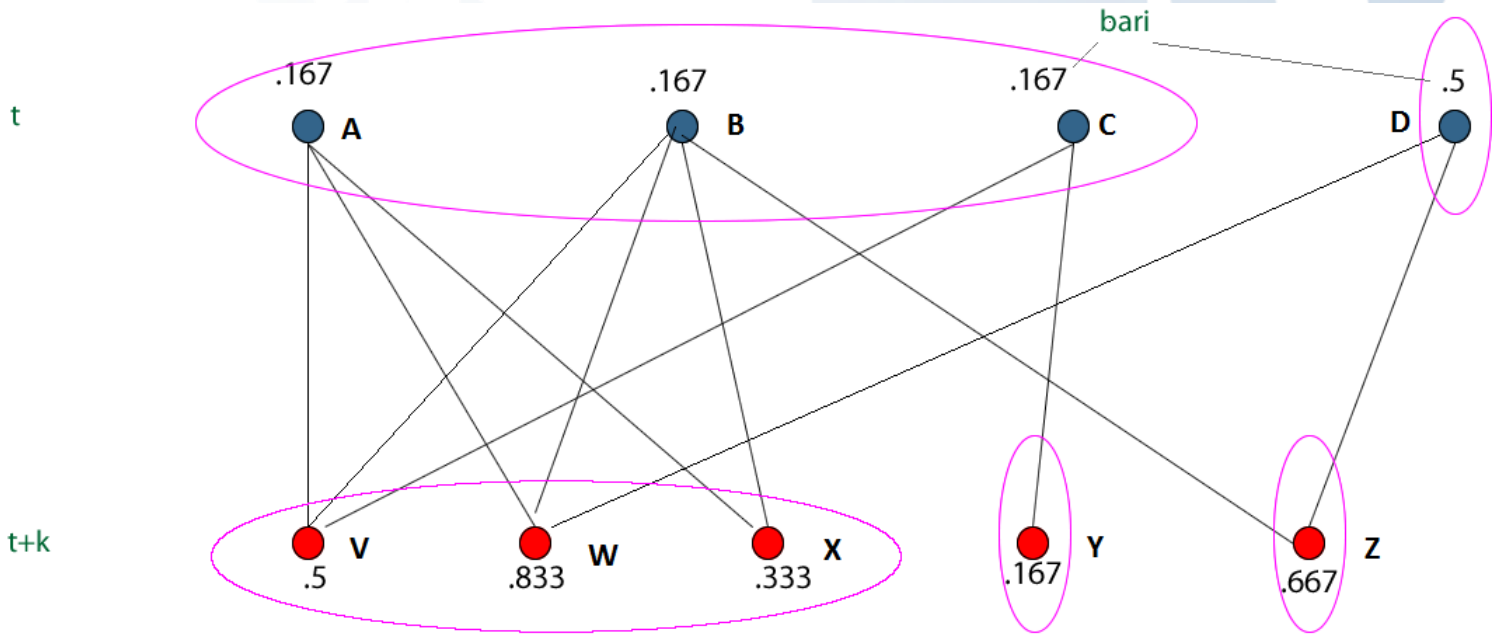
- Need to turn 1974 antecedents of 1996 into a representative sample.
- General issue turns out to be how trace forward and trace backward of panels are collected
- In initial round of REDS the followed up households from ARIS were only sampled if the head had not died and the household had not split.
- In MHSS and many other such surveys in low-income countries follow-up of all household members. This maximizes recontact rates of individuals In other surveys focus is specifically on descendants
- Each of these has different implications for how you would construct design-based weights in a cross-section

Indirect Sampling

- Problem is known as indirect sampling–a sample of descendants (or antecedents) of an initial population-based sample
- Horovitz-Thompson estimate–requires knowledge of overall probability of being in indirect sample, which in general requires knowledge of joint distribution of sampling for all antecedents–not just the particular antecedent that resulted in the inclusion of that household.
  $\mathbb{E} \sum y_i \frac{I_i}{\pi_i} = \sum y_i \frac{\mathbb{E} I_i}{\pi_i} = \sum y_i$
- The generalized weight share method works if at least one antecedent j is observed.
  $\mathbb{E} \sum y_i \sum_{j \in A_i} \gamma_j \frac{I_j}{\pi_j} = \sum y_i \sum_{j \in A_i} \gamma_j \frac{\mathbb{E} I_j}{\pi_j} = \sum y_i$ where $\sum \gamma_i = 1$.
  Note that this requires knowledge of the number of antecedents.

Illustration

Figure 1: Sampling Probabilities with Household Recombination



Like Marries Like Solution I

- The standard approach for large national panels like the PSID–and a version of the weight-share method.
- Each person has a mother and father, with respective sampling probability $p = \tilde{p}k$ and $q = \tilde{q}k$ where k is inversely proportional to population size.
- Sample weight w scaled by k assuming independence is $wk = \frac{1}{1-(1-p)(1-q)}k = (\tilde{q}+\tilde{p})^{-1} + \frac{\tilde{p}\,\tilde{q}}{(\tilde{q}+\tilde{p})^2}\,k + O\left(k^2\right)$
- Only observe sampling parent for the parent who is in the sample
- Assume sampling probability of unsampled individual is the same as sampled individual (Like Marries Like).
- So sampling weight when the mother is from the panel is $\frac{1}{1-(1-p)^2}$, which is likely to be wrong in practice.

Like Marries Like Solution II

- But sometimes only father is from the sample and sometimes it is both so

$$\mathbb{E}(wk) = \frac{k}{1-(1-p)(1-q)}\left(\frac{p(1-q)}{1-(1-p)^2} + \frac{q(1-p)}{1-(1-q)^2} + \frac{pq}{1-(1-p)(1-q)}\right)$$
$$= (\tilde{q}+\tilde{p})^{-1} + \frac{-\tilde{p}^2+6\,\tilde{p}\,\tilde{q}-\tilde{q}^2}{4\,(\tilde{q}+\tilde{p})^2}k + O\left(k^2\right)$$

- Compare to actual
$= (\tilde{q}+\tilde{p})^{-1} + \frac{\tilde{p}\,\tilde{q}}{(\tilde{q}+\tilde{p})^2}\,k + O\left(k^2\right)$
- Efficiency can be improved if account for known differences in sampling probabilities within spouse (e.g., interracial spouses–Foster and Pellerin 2019).
- But our sample, even if it were based on biology rather than coresidence (a) goes backwards (we all have two direct antecedents but not necessarily exactly two descendants), (b) is not necessarily independent (bari sampling), (c) sample is not small (10-20 percent), and (d) 1/3 parents from outside of the frame.

Definition of Descendant Household

- Based on residence rather than relationships
- Someone in the 1993/2014 household also lived in the 1974 household (Zero Order Link)
- Someone in the 1993/2014 household has lived with a member of the 1974 household at any point between 1974 and 1993/2014 (First Order Lived with Link)
- Someone in the 1993/2014 household has lived with a person who lived with a member of the 1974 household prior to living with the 1993/2014 person (Second Order Forward Lived with Link)

Household Links Construction I

- H_1 : Partitions people (a, b, c, d) into households (α, β) at time 1
- H_2 : Partitions people (b, c, d, e, f) into households $(\delta, \gamma, \epsilon)$ at time 2
- P_1 : Maps people from period 1 to period 2

$$H_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cc} \alpha & \beta \end{array} \\ \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \end{array} \qquad H_2 = \begin{array}{c} \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{ccc} \delta & \gamma & \epsilon \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

$$P_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{ccccc} b & c & d & e & f \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Household Links Construction II

- $C_t = H_t * H_t^T$: Who is coresident with whom at time t
- $LI_1 = C_1 * P_1 * C_2$: Links of people at time 1 to people at time 2 based on coresidence in each period
- $LH_1 = H_1^T * LI_1 * LI_2 * ... * LI_{t-1} * H_t$: Links of households at time 1 to households at time t

$$C_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

$$LI_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{ccccc} b & c & d & e & f \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

$$LH_1 = \begin{array}{c} \\ \alpha \\ \beta \end{array} \begin{array}{c} \begin{array}{ccc} \delta & \gamma & \epsilon \end{array} \\ \begin{pmatrix} 2 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \end{array}$$

Links in Practice

Table 1: Number of Links Between Selected 1974 and 1996 Households

	1996 Households										
	A11*	B11	C11*	D11	D12	D13	E11	C12	F11	B12	B13
1974 Households											
A01	3	4	0	0	0	0	0	0	0	0	0
D01	3	4	7	5	4	4	7	4	5	1	4
G02	3	4	0	0	0	0	0	0	0	0	0
D02	0	0	7	5	4	0	0	4	5	3	4
D03	0	0	1	3	3	0	0	0	5	0	0
C01	0	0	7	0	0	0	0	4	0	0	0
C02	0	0	7	0	0	0	0	4	0	0	0
E01	0	0	0	0	0	1	7	0	0	0	0
H01	0	0	0	0	0	0	0	0	0	1	2
J01	0	0	0	0	0	0	0	0	0	3	0
K01	0	0	0	0	0	0	0	0	0	3	4
G01	0	0	0	0	0	0	0	0	0	3	4
L01	0	0	0	0	0	0	0	0	0	3	0
B01	0	0	0	0	0	0	0	0	0	1	2

*Household was in the 1996 MHSS

Constructing probabilities

- Note original sample is 1996 and indirect sampling is for 1974 (and 2012).
- Key is survey is inside a Demographic Surveillance System, which also provided the sampling frame.
- Can reconstruct households at each point in time and link individuals across time.
- So redraw 1996 sample many times and each time identify antecedents.
- Average fraction of times 1974 household appears in samples is the sampling probability.

Sampling on outcome

```
sampling.log                                                    7/20/2019

. set obs 100000
. gen e=invnorm(uniform())
. gen x=uniform()
. gen y=x+e
. gen ps=exp(y)/(1+exp(y))
. reg y ps

      Source |       SS           df       MS      Number of obs   =   100,000
-------------+----------------------------------   F(1, 99998)     >  99999.00
       Model |  106209.486         1  106209.486   Prob > F        =    0.0000
    Residual |  2885.91114    99,998  .028859689   R-squared       =    0.9735
-------------+----------------------------------   Adj R-squared   =    0.9735
       Total |  109095.397    99,999  1.09096488   Root MSE        =    .16988

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ps |     4.94921   .0025799  1918.39   0.000     4.944154    4.954267
        cons |  -2.473878   .0016397 -1508.75   0.000    -2.477092   -2.470664
------------------------------------------------------------------------------
```

Sampling on outcome

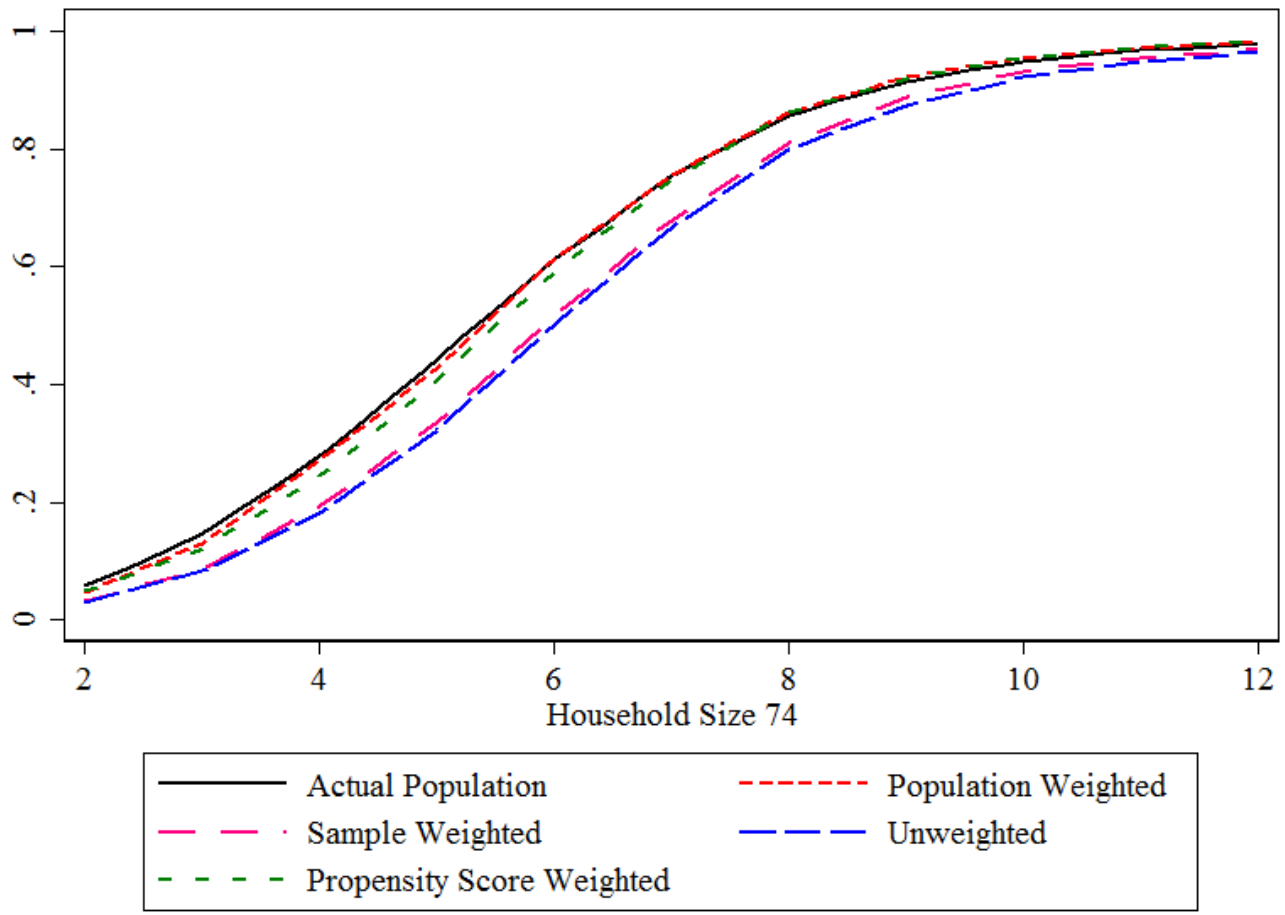
```
. gen sample=uniform()<ps
. gen wt=1/ps
. reg y x if sample

      Source |       SS           df       MS      Number of obs   =    59,688
-------------+----------------------------------   F(1, 59686)     =   4298.80
       Model |  3644.70609         1  3644.70609   Prob > F        =    0.0000
    Residual |  50604.3012    59,686  .847842061   R-squared       =    0.0672
-------------+----------------------------------   Adj R-squared   =    0.0672
       Total |  54249.0073    59,687  .908891506   Root MSE        =    .92078

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .8612113   .0131352    65.57   0.000     .8354663    .8869563
        cons |   .4028977    .007891    51.06   0.000     .3874312    .4183642
------------------------------------------------------------------------------

. reg y x if sample [aw=wt]
(sum of wgt is 99,457.7167572975)

      Source |       SS           df       MS      Number of obs   =    59,688
-------------+----------------------------------   F(1, 59686)     =   5326.42
       Model |  5378.12395         1  5378.12395   Prob > F        =    0.0000
    Residual |  60265.4173    59,686  1.00970776   R-squared       =    0.0819
-------------+----------------------------------   Adj R-squared   =    0.0819
       Total |  65643.5412    59,687  1.09979629   Root MSE        =    1.0048

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.039937   .0142492    72.98   0.000     1.012009    1.067866
        cons |    -.02221   .0082187    -2.70   0.007    -.0383186   -.0061014
------------------------------------------------------------------------------
```

Sampling on outcome

- Note that e_i is correlated with π_i but expectation is only with respect to sampling
- $\mathbb{E}\,\frac{1}{N}\sum x_i e_i \frac{I_i}{\pi_i} = \frac{1}{N}\sum (x_i e_i)\frac{\mathbb{E}\,I_i}{\pi_i} = \frac{1}{N}\sum x_i e_i$
- So unbiased as long as unbiased in population
- Fits within general literature on choice-based sampling.

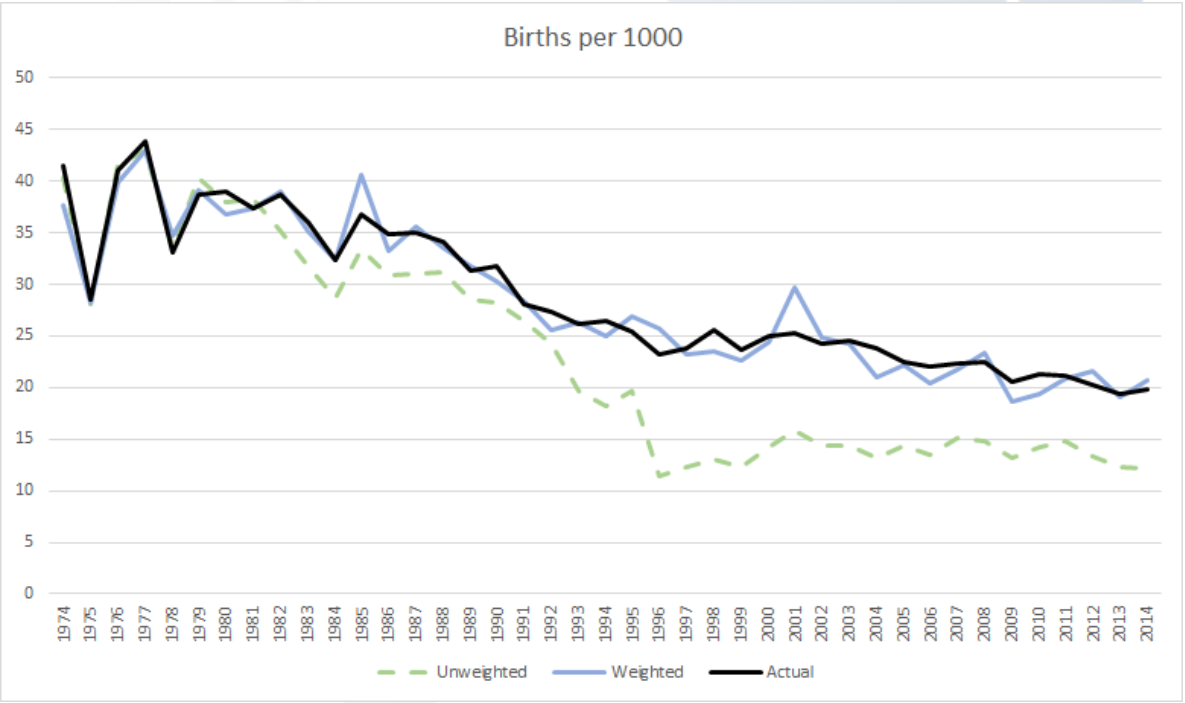
Does it matter?

Figure 2: Estimated Distribution of 74 Household Size with Various Weights



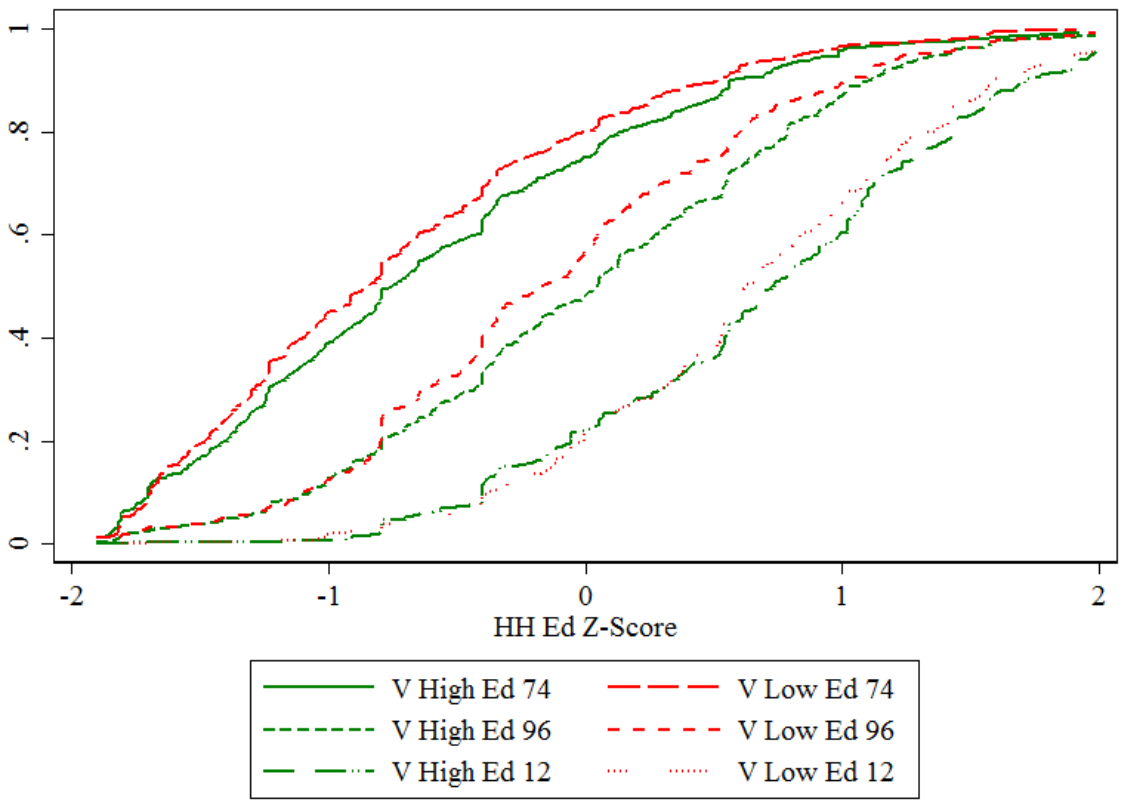
Population, Weighted and Unweighted Estimates

Figure 3: Births



Cross-sectional change

Figure 4: Distribution of HH Ed Z-score by Year and Village Ed



Mobility

- The presence of panel data over 40+ years in principle allows us to look at economic mobility.
- To what extent are the gains in access to education/health/labor market opportunities concentrated among well-off households
- Is the degree of economic mobility different in areas with early access to MCH-FP services?

What population are we interested in? I

- Population of individuals
  - Population is clear
  - Limited use over long periods of time
  - Not suitable for looking at changes in age-specific behavior like schooling
  - Individual well-being and its measurement affected by household coresidence

What population are we interested in? II

- Population of biological descendants
  - Clearly defined in principle
  - Biological descendants are often only visible if there is coresidence
  - We do not typically have data on all relevant biological antecedents (e.g., spouse's origin)
  - Generations may be asyncronous (e.g., set of grandchildren born over a large span of time)

What population are we interested in? III

- Population defined by households in region
  - Well-defined at each point in time
  - Political and administrative units are spatially defined
  - Symmetry between antecedant and descendent populations
  - Can be implemented in a regionally defined surveilance system
  - Misses consequences of programs, for example, for outmigrants

Nonrandom sampling of descendants

- In a regular longitudinal sample one would follow all or a random subset of descendants
- But in retrospective evaluation, "followed" households are a random subsample of population but not necessarily of descendants.
- Given bari-sampling descendant households from the same 1974 antecedent in the same bari will never both show up in the sample
- Related problem arises if following individuals because coresident members are selected together.
- Same problem arises in a regular longitudinal sample if interested in characteristics of antecedents

Example I

- Suppose two descendant households. If we sample both average income is $\frac{y_1+y_2}{2}$
- If independent sampling then define $\bar{y} = y_i$ if only y_i is observed and $\bar{y} = \frac{y_1+y_2}{2}$ if both are observed, so $E\bar{y} = \frac{y_1+y_2}{2}$ is unbiased.
- But suppose state $B = \{y_1, y_2\}$ is observed with probability 1/2 and state $C = \{y_2\}$ with probability 1/2.
- Taking expectations across states $E\bar{y} = 1/4y_1 + 3/4y_2$ is biased.
- Two options: (a) throw out y_2 if state B (b) throw out observation if state C
- Both seem wasteful and how does this generalize if $\{y_1, y_2\}$ are observed with probability 1/3?

Alternate Measures I

For a 1974 hhold j with 2 descendants of which at least one is picked:

- option A=hhold 1 picked
- option B=hholds 1 and 2 picked
- option C=hhold 2 picked

Find weights w_a, w_b, w_{b1}, w_{b2}, and w_c such that:

$$\mathbb{E}(y) = p_a w_a y_1 + p_b w_b (w_{b1} y_1 + w_{b2} y_2) + p_c w_c y_2 = \frac{1}{2} y_1 + \frac{1}{2} y_2 = \bar{y} \quad (1)$$

Alternate Measures II

We also want to minimize the effect that the variation in the fraction of households in each sample has on y
We therefore want to minimize:

$$\begin{aligned} Z = [&var(p_a)(w_a^2 y_1^2) + var(p_b)(w_b^2)(w_{b1}y_1 + w_{b2}y_2)^2 \\ &+ var(p_c)(w_c^2 y_2^2) - 2cov(p_a, p_b)(w_a y_1)(w_{b1}y_1 + w_{b2}y_2) \\ &- 2cov(p_a, p_c)(w_a y_1)(w_c y_2) - 2cov(p_b, p_c)(w_{b1}y_1 + w_{b2}y_2)(w_c y_2)] \end{aligned} \quad (2)$$

Minimization Problem Continued

Taking derivatives of (1), we get the following two conditions:

$$p_a w_a + p_b w_b w_{b1} = \frac{1}{2} \tag{3}$$
$$p_b w_b w_{b2} + p_c w_c = \frac{1}{2}$$

The criterion function we then use is the sum of the two second derivatives of equation (2):

$$\min_{w_a, w_b, w_{b1}, w_{b2}, w_c} \frac{d^2 Z}{dy_1^2} + \frac{d^2 Z}{dy_2^2} \tag{4}$$

To find the best possible weights without knowing income, we minimize equation (4) subject to equations (3)

Solution to Minimization Problem

Solving the above minimization problem we find that the weights which minimize the variance are based on the probability of sampling a 1996 hhold

In our two household example, we get the following weights:

$$w_a = w_b w_{b1} = \frac{\Pr(j)}{2 * \Pr(1)} = w_1$$

$$w_c = w_b w_{b2} = \frac{\Pr(j)}{2 * \Pr(2)} = w_2$$

We can generalize this result to assign a weight to every descendant i of a 1974 household j with N descendants:

$$w_i = \frac{\Pr(j)}{N * \Pr(i)}$$

This is the weight-share method. It requires knowledge of number of descendants in addition to survey design probabilities.

Formal Definition of Mobility Weights

- We wish to estimate $\Delta \bar{c}_{tN} = \frac{1}{N_{tx}} \sum_{i \in I_{tx}} \frac{1}{|A^{-1}(i)|} \sum_{j \in A^{-1}(i)} \left(c_{jt+1} - c_{it} \right)$
- and $\Delta \hat{c}_{tN} = \frac{1}{N_{tx}} \sum_{i \in I_{tx}} \frac{1}{|A^{-1}(i)|} \sum_{j \in A^{-1}(i)} \left(c_{jt+1} - c_{it} \right) \frac{\mathbb{1}(j \in S_{t+1})}{\mathbb{E}(\mathbb{1}(j \in S_{t+1}))}$
- then $\operatorname*{plim}_{N \to \infty} \frac{\Delta \hat{c}_{tN}}{\hat{i}_{tN}} - \Delta \bar{c}_{tN} = 0.$

Applying Formalism

- First apply weights to descendants
- Weighted values are averaged across descendants
- 1974 cross-sectional weights are then applied

Simulation i

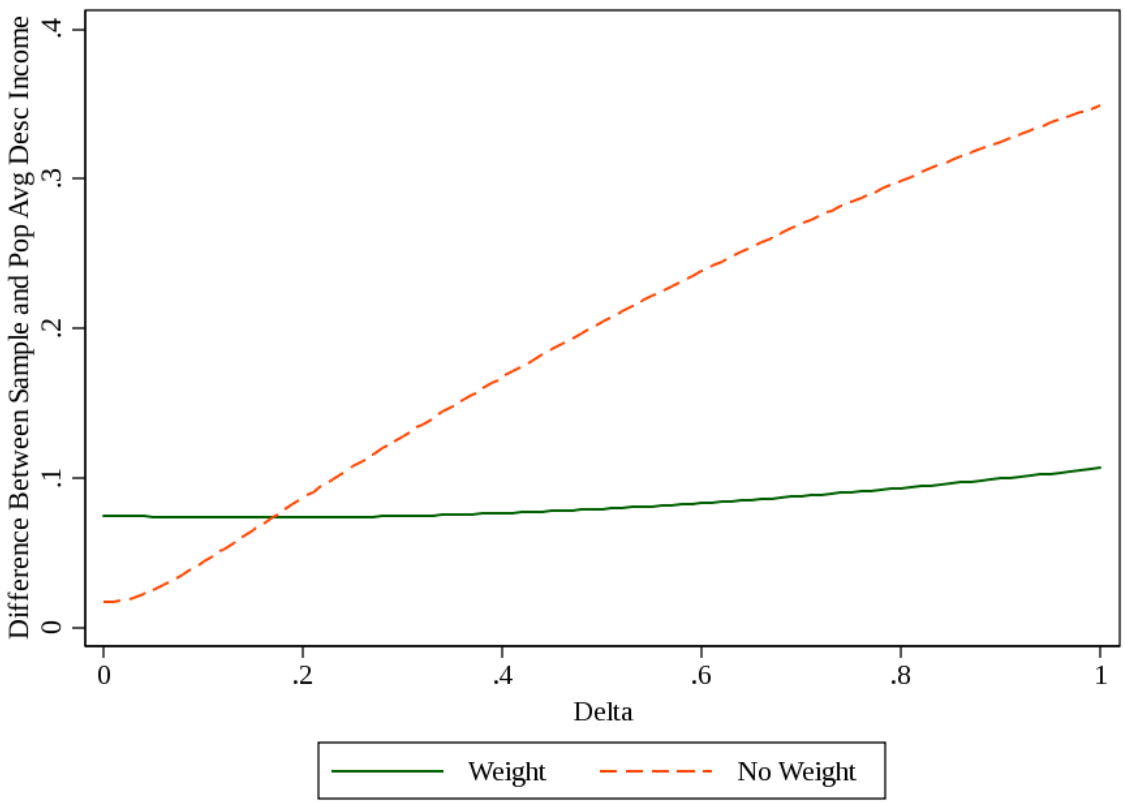
- Simulation exercise with 3 states of the world $\{y_1, y_2, \{y_1, y_2\}\}$ and the probability of each state of the world is correlated with the average y of those households observed.
- Parameter δ measures strength of the correlation

Simulation II

$$
\begin{aligned}
Pr(1) &= \frac{e^{\delta * y_1}}{e^{\delta * y_1} + e^{\delta * y_2} + e^{\delta * \bar{y}}} \\
Pr(2) &= \frac{e^{\delta * y_2}}{e^{\delta * y_1} + e^{\delta * y_2} + e^{\delta * \bar{y}}} \\
Pr(1\&2) &= \frac{e^{\delta * \bar{y}}}{e^{\delta * y_1} + e^{\delta * y_2} + e^{\delta * \bar{y}}}
\end{aligned}
\tag{5}
$$

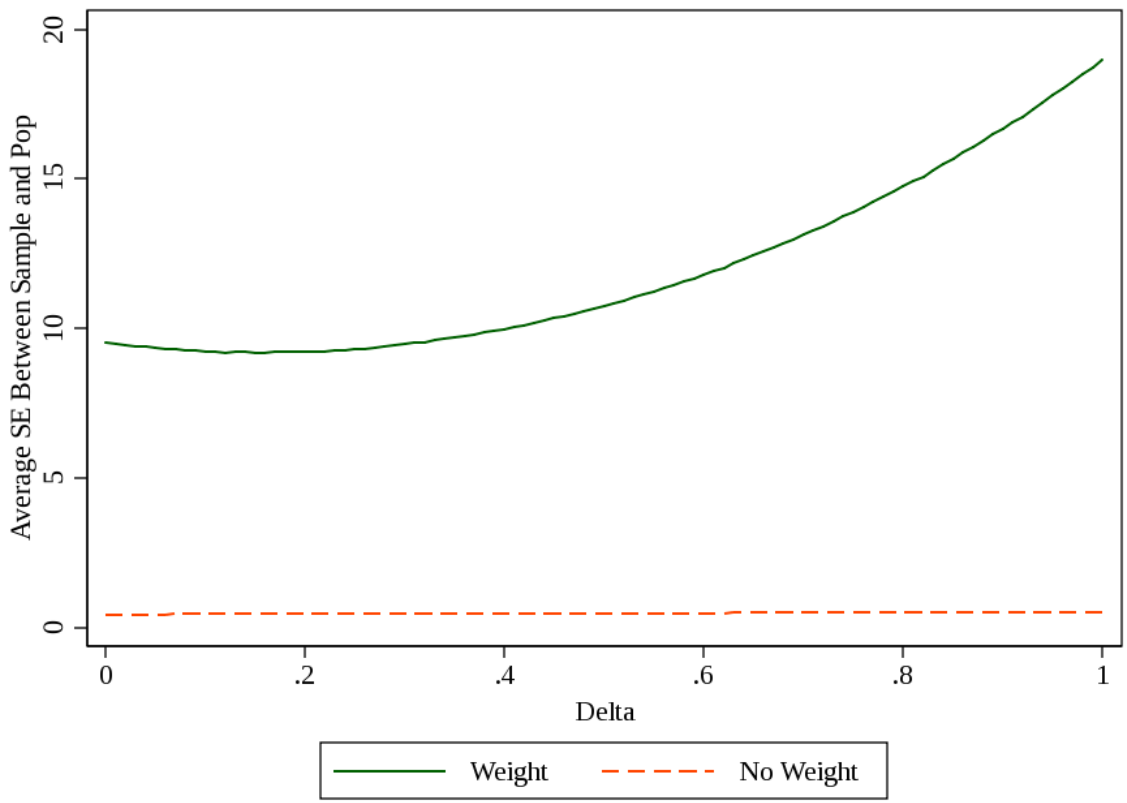
Simulation III

Figure 5: Average absolute difference between sample and actual descendant income means for different levels of correlation



Simulation IV

Figure 6: Average squared difference between sample and actual descendant incomes



How well does it work?

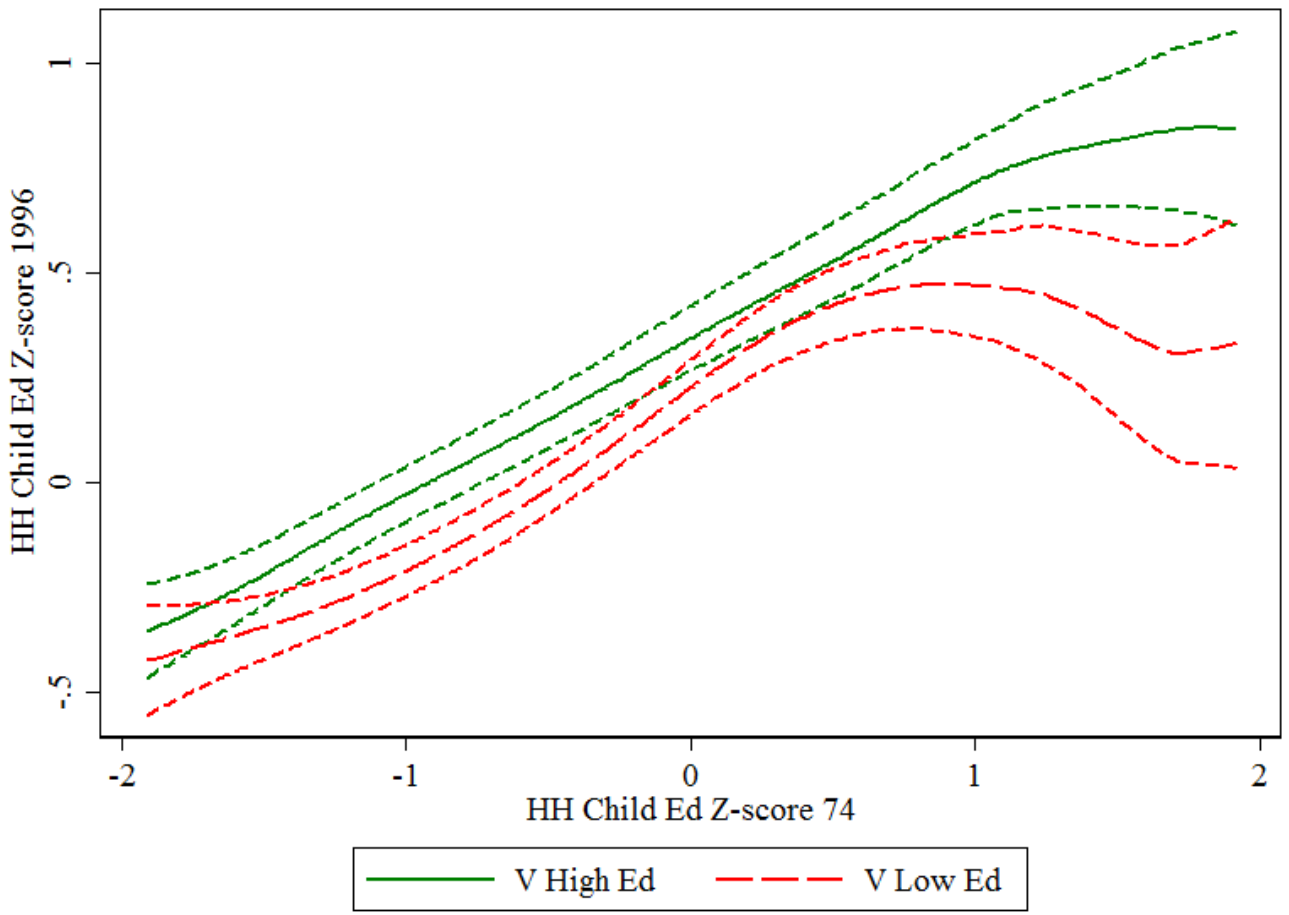
Table 2: Household Size Change by 74 Conditions and 74 Village Ed

VARIABLES	(1) Population	(2) Formal	(3) Predicted	(4) 74 Weights	(5) 74/96 Weights	(6) 96 Weights	(7) No Weights
Ed Low	0.0223	0.018	0.071	0.022	0.024	0.003	-0.002
	(0.0344)	(0.155)	(0.347)	(0.140)	(0.140)	(0.088)	(0.086)
Ed High	-0.110	-0.111	0.345	-0.101	-0.098	-0.167	-0.172
	(0.0433)	(0.184)	(0.381)	(0.167)	(0.168)	(0.112)	(0.110)
H Size Low	2.046	2.057	3.439	1.996	2.005	2.058	2.043
	(0.0335)	(0.148)	(0.352)	(0.138)	(0.138)	(0.085)	(0.084)
H Size High	-3.411	-3.414	-4.669	-3.374	-3.384	-3.608	-3.594
	(0.0526)	(0.229)	(0.457)	(0.187)	(0.187)	(0.105)	(0.103)
Cons Low	-0.597	-0.593	-0.727	-0.590	-0.593	-0.683	-0.679
	(0.0391)	(0.173)	(0.347)	(0.157)	(0.157)	(0.091)	(0.089)
Cons High	0.303	0.305	0.485	0.307	0.309	0.314	0.308
	(0.0394)	(0.176)	(0.391)	(0.164)	(0.164)	(0.098)	(0.096)
V High Ed	-0.181	-0.178	-0.209	-0.175	-0.174	-0.156	-0.156
	(0.0320)	(0.142)	(0.298)	(0.130)	(0.130)	(0.081)	(0.080)
Constant	-0.803	-0.811	-1.901	-0.683	-0.701	-0.731	-0.692
	(0.0418)	(0.184)	(0.401)	(0.169)	(0.170)	(0.119)	(0.118)
Observations	19,820	4,690	4,688	4,690	4,690	4,690	4,690
R-squared	0.310	0.261	0.159	0.325	0.328	0.339	0.335

Standard deviation in parentheses for columns 2-7
Standard error in parentheses for column 1

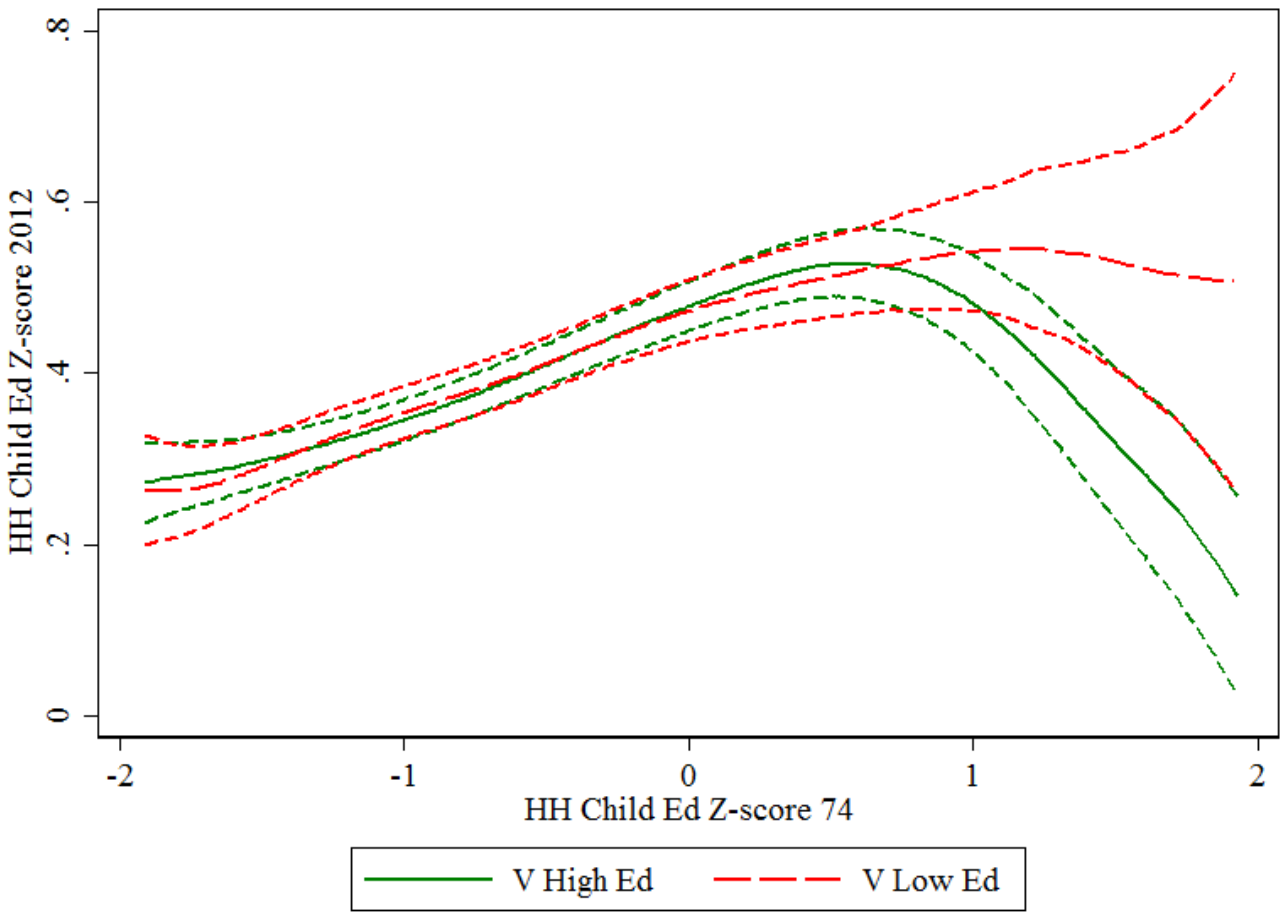
1996 Educational Mobility Graph

Figure 7: HH Education by 74 HH Education and 74 Village Ed



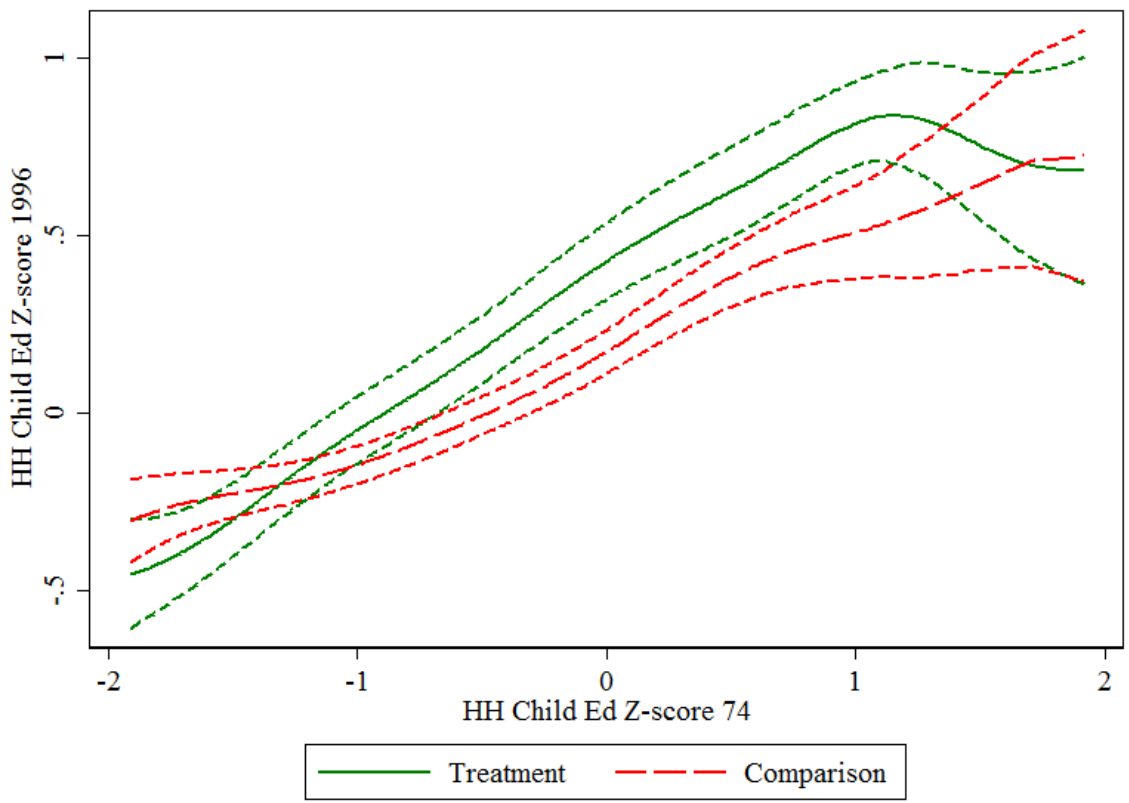
2012 Educational Mobility Graph

Figure 8: HH Education by 74 HH Education and 74 Village Ed



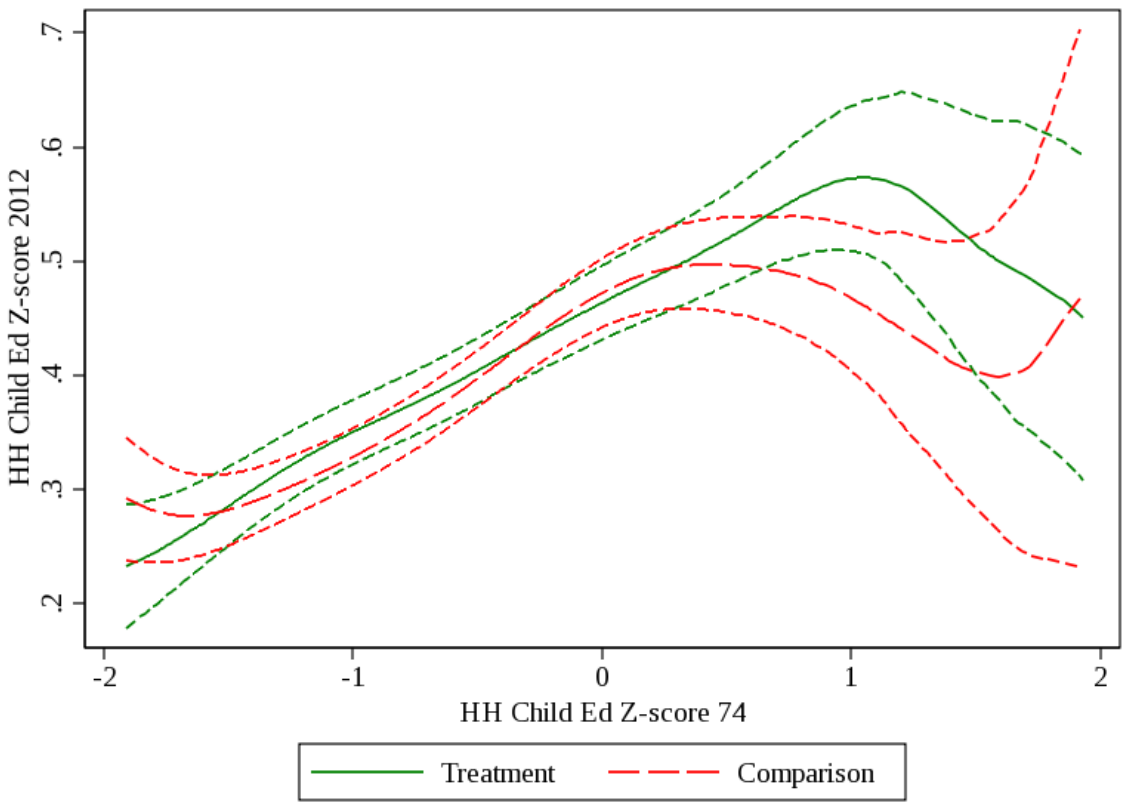
Treatment Comparison Mobility 1996

Figure 9: Education



Treatment Comparison Mobility 2012

Figure 10: Education



Income Segregation

Logan, Foster, Ke, and Li (2018)

- Several studies have found that income segregation in urban areas of the US is rising
  - Especially since 2000
  - Especially among minority sub-populations
  - Up to 2000 data are from Census Long-Form and after 2005 are from the American Community Survey.
  - Sampling rates: Long-From average 15% and ACS average 5%
- How does the change in sampling affect measurement of income segregation?
- Do conclusions about segregation change need to be modified in light of this concern?

Segregation Measures

- NSI Neighborhood Sorting Index: $\left(\frac{\sum(\bar{y}_i-\bar{y})^2}{\sum(y_ij-\bar{y})^2}\right)^{1/2} = \left(\frac{AV}{TV}\right)^{1/2}$
- H Rank Order Information Theory Index, based on $E_i(q) = p_i(q)\ln(p_i(q)) + (1-p_i(q))\ln(1-p_i(q))$ where p_i is the fraction of households in tract i with income less than q, then integrated over q for city as a whole and added up across tracts.
- H90 Entropy Index for $q = 90$ (measures segregation of richest households)
- H10 Entropy Index for $q = 10$ (measures segregation of poorest households)
- R, R90, and R10 Rank Order Variance Ratio Indices

What goes wrong?

- Take NSI
  - If all tracts have the same distribution and tract variances are non-zero then for the population NSI=0. There is no segregation.
  - But if we just sample one household per tract there will be variance across tracts in the tract "means" so for the sampled households NSI>0
  - As we sample more households the within tract mean approaches the true mean so the bias in the NSI vanishes.

NSI bias correction

- AV, which is based on tract means, is not well-measured when sampled units per tact is small
- but $AV + WV = TV$
- TV and WV are well estimated if number of tracts is large–so we can estimate AV in a roundabout way.
- How is that tract variances are well-measured when tract means are not?
- Unbiased but noisy estimate of WV for tract i is $\hat{\sigma_i^2} = \frac{1}{N_i - 1}(y_{ji} - \bar{y}_i)^2$
- $\mathbb{E}\,\hat{\sigma_i^2} = \sigma_i^2$
- $\frac{1}{T}\sum \hat{\sigma_i^2} \xrightarrow{p} \frac{1}{T}\sum \sigma_i^2$
- In principle this works even if just sample two households per tract as long as there are enough tracts as $\mathbb{E}\,\frac{1}{N-1}((y_{i1} - \bar{y})^2 + (y_{i1} - \bar{y})^2) = \mathbb{E}\,\frac{1}{2}(y_{i1} - y_{i2})^2 = \sigma_i^2$
- Call this Small Sample Variance Decomposition (SSVD)

Correcting Entropy I

- SSVD works for any quadratic based index such as the rank-order variance ratios: R, R90, and R10.
- It does not work for entropy based measures
- Suppose p is the proportion of households below some income in the tract population and s the proportion of households below some income in the tract sample
- If $h(s) = sln(s) + (1-s)ln(1-s)$ is the entropy function
- then the bias is $\mathbb{E}((h(s) - h(p))|p)$ where expectation is across possible samples.
- But the distribution of s depends on p and if you know p you would have already solved the bias problem!

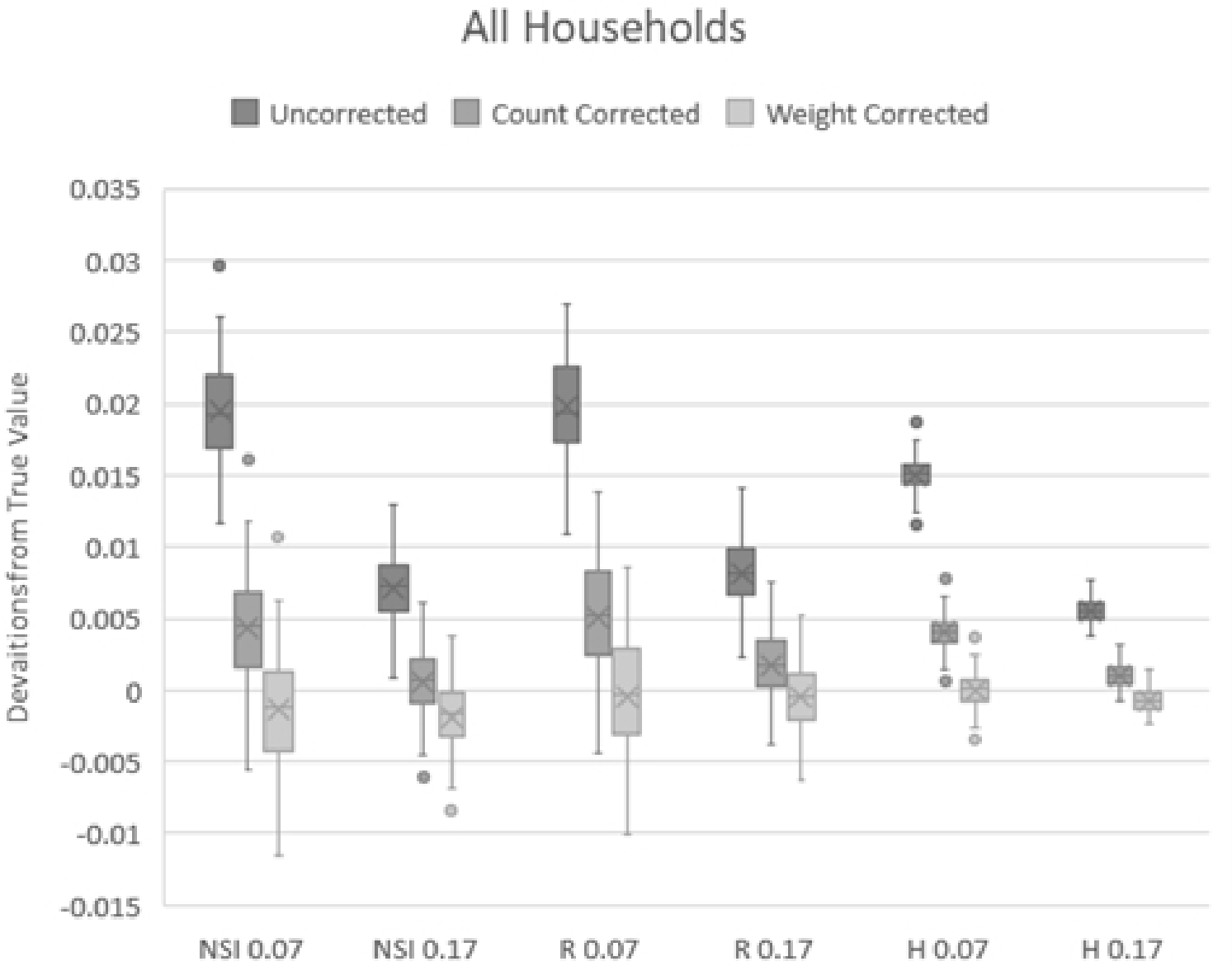
Correcting Entropy II

- But if the number of sampled households per tract N is reasonably large then s is not far from p so we can carry out a second-order Taylor approximation around $s = p$:
- So
  $\mathbb{E}\, h(s) \approx h(p) + (ln(p) + ln(1-p))\, \mathbb{E}(p-s) + \frac{1}{2p(1-p)}\, \mathbb{E}((p-s)^2)$
- but $\mathbb{E}(p-s) = 0$ and $\mathbb{E}((p-s)^2) = \frac{1}{N}\frac{M-N}{M-1}p(1-p)$
- so $\mathbb{E}\, h(s) \approx h(p) + \frac{1}{N}\frac{M-N}{M-1}$
- or $\frac{1}{N}$ if the population tract size is large (Miller 1955)
- So bias depends essentially on the harmonic mean of the tract sample sizes.

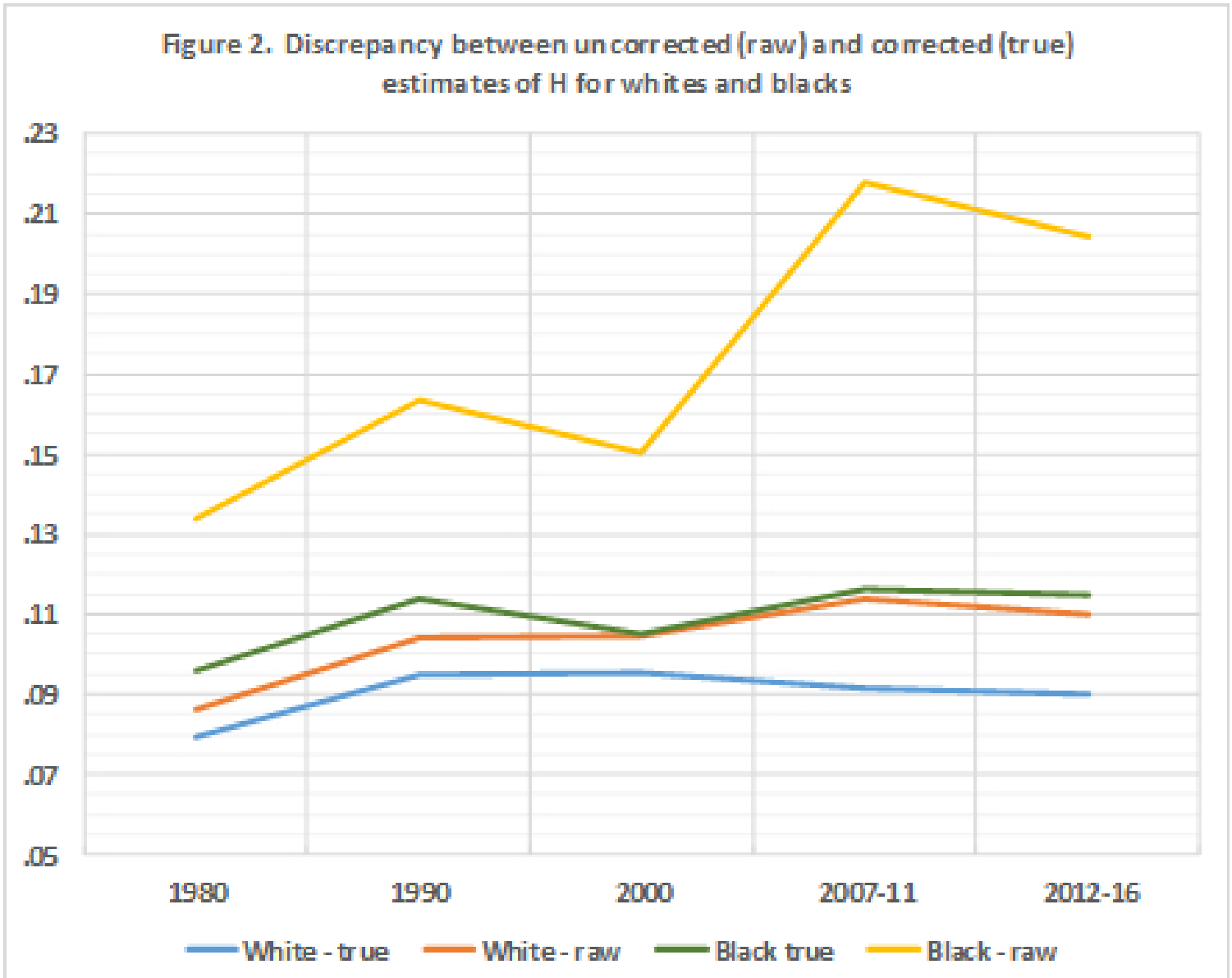
Adding weights

- Ran simulations from 1940 manuscript to test and then applied to grouped data from 2000-2010.
- But in RDC recognized the importance of weights in construction of grouped data
- So derived approximate bias using weights
- $H_{gw} - H_{bw} = -\sum_j \frac{M_j}{M} \sum_i {w_{ij}}^2$
- $WI = \sum_j \frac{M_j}{M} \frac{1}{1-\sum_i {w_{ij}}^2} \sum_i w_{ij}(y_{ij} - \bar{y}_j)^2$
- $TO = \sum_j \frac{M_j}{M} \sum_i (w_{ij} y_{ij} - \bar{y})^2$
- $NSI_{gw} = \left(1 - \frac{WI}{TO}\right)^{1/2}$

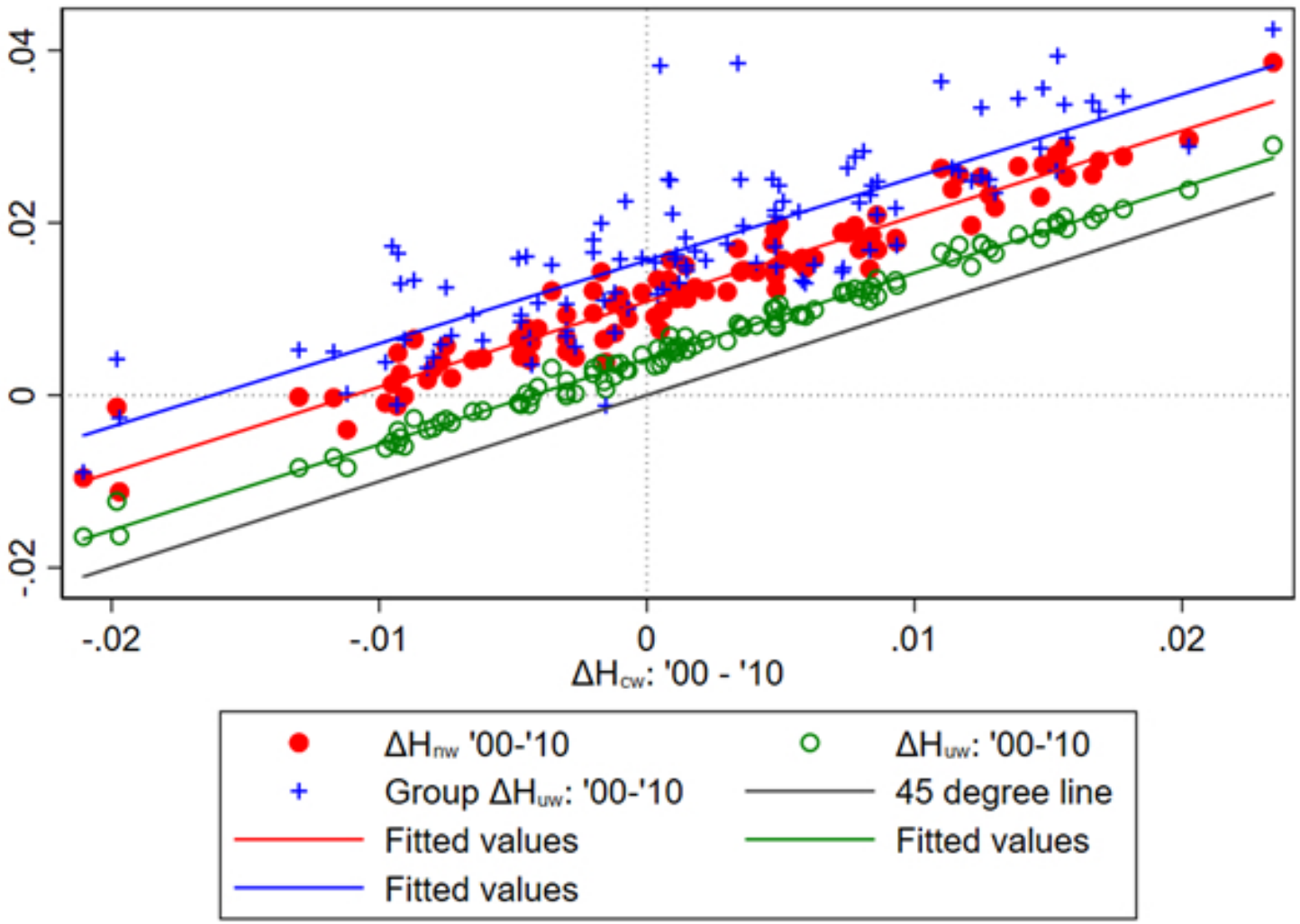
1940 Simulations



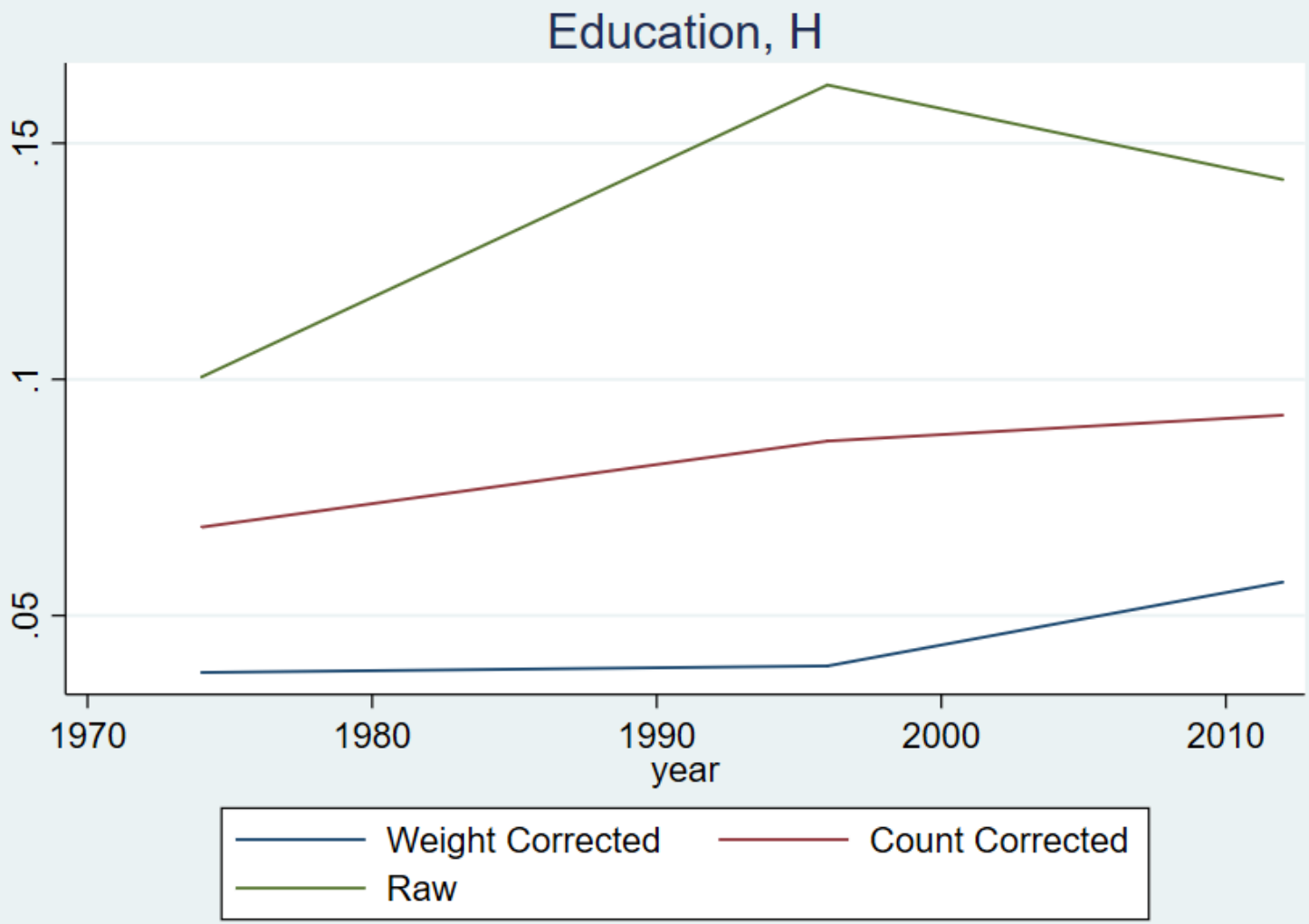
Actual Average Trends



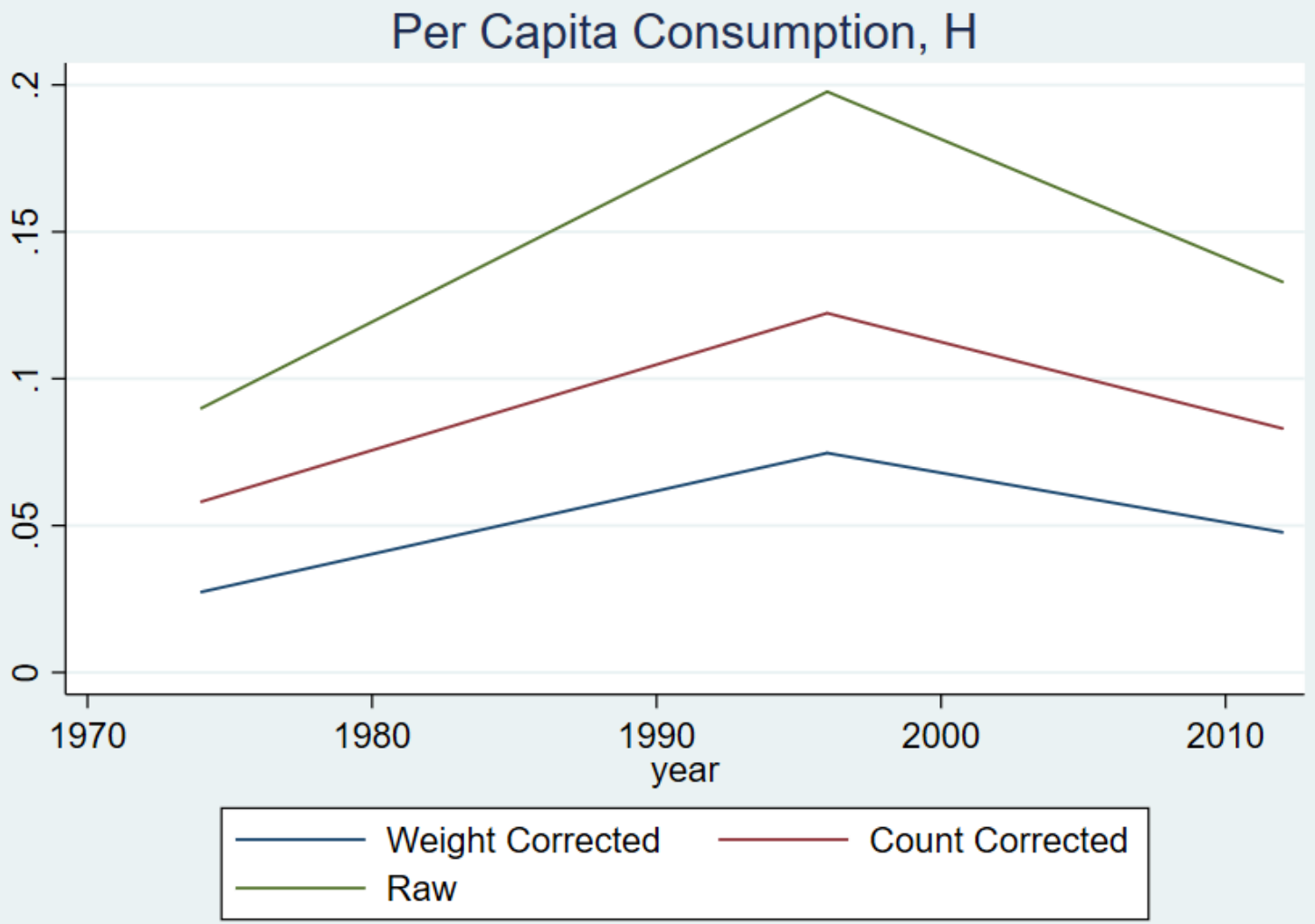
Effects of Different Elements



Village Educational Segregation Matlab H



Village Per Capita Consumption Segregation Matlab, H



Effects of Institutions

- Dalbo, Foster, Putterman (2010) devised an experiment to see if a policy implemented democratically led to a different result than the same policy implemented randomly, controlling for the selection into the policy under democracy.
- Key insight was that conditional on vote unobservables were uncorrelated with policy choice.
- Method required voting even in random case. This may be difficult to implement outside the lab and may have direct effects.
- Dalbo, Foster, Kamei (2018) showed that could get same result without voting through a weighting procedure.
- Weight choices conditional on policy by share of votes in the population.

Reweighting

- Average Cooperative behavior as a function of democracy(D),non(N) and policy E.
  $\mathbb{E}\, C(N,E) = P(v=a|N,E)\int C(N,E,\mu)dF(\mu|a) + P(v=b|N,E)\int C(N,e,\mu)dF(\mu|b)$
- $\mathbb{E}\, C(D,E) = P(v=a|D,E)\int C(D,E,\mu)dF(\mu|a) + P(v=b|D,E)\int C(D,e,\mu)dF(\mu|b)$
- These can differ because voting is correlated with policy or because cooperation is different by regime
- But voting is not correlated with policy under non. $\mathbb{E}\, C(N,E) = P(v=a)\int C(N,E,\mu)dF(\mu|a) + P(v=b)\int C(N,e,\mu)dF(\mu|b)$
- And can reweight by overall voting under democracy.
  $\mathbb{E}\, C_{RW}(D,E) = P(v=a)\int C(D,E,\mu)dF(\mu|a) + P(v=b)\int C(D,e,\mu)dF(\mu|b)$

Payoffs

Table 2: Stage Games - Dal Bó, Foster and Putterman (2010)

Initial payoffs			Modified payoffs		
	Other's action			Other's action	
Own action	C	D	Own action	C	D
C	50	10	C	50	10
D	60	40	D	48	40

Weighting Adjustment

Table 3: The effect of the democracy in overcoming social dilemmas

	Mechanism						
	Democracy				Random	Democracy Effect	Standard Errors[p-values]
Environment	Voted Yes	Voted No	All	Properly weighted average			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Initial	24	14.55	17.5	19.57	18.06	1.51	6.7
	(25)	(55)	(80)		(72)		[0.8215]
Modified	80	40	70	61.25	47.92	13.33	6.48
	(60)	(20)	(80)		(192)		[0.0396]**

Notes: Data from Dal Bó, Foster and Putterman (2010). Numbers in parentheses indicate the number of subjects. Bootstrapped SE. The number of bootstrap iterations is 10,000. ** indicates significance at the 0.05 level.

Conclusions

- Distinction between idiosyncratic shocks that affect the process of household formation and dissolution and the "pure" randomness that comes from the sampling process.
- Constructing appropriate weights may be hard in panel data without equivalence of surveillance data–can we collect retrospective information that helps?
- Follow up strategy has implications for both construction of cross-sectional weights and for measurement of economic mobility.
- Sample sizes and weights matter for measuring spatial heterogeneity using survey data
- Weights can adjust for certain types of selectivity bias.