A User-Friendly Technique for Implementing Survey Weights Using Stata

Islam Rabiul[1] Sweetman Arthur[2]

[1]Ph.D. Candidate
Department of Economics
McMaster University

[2]Professor
Department of Economics
McMaster University

Canadian Stata Conference, September 23, 2021

Table of Contents

Motivation

- Survey data frequently suffer from bias for many reasons
  - Non-random non-response
  - Non-random attrition
  - Non-random solicitation of respondents
- Survey statisticians recommend using survey weights to improve quality of estimation when selection on observables & if "sufficient" auxiliary data available
- `Auxiliary` data available from a large sample or population
  - Data that come from out of sample to calibrate with are known as `Auxiliary` data

Motivation

- **Reintroducing:** An easy to use GMM method of weighted regression analysis using auxiliary data by Imbens & Lancaster, and Hellerstein and Imbens
  - Developing Stata program
  - Easy to implement in Stata
- This method can be used for wide variety of estimators (any GMM estimators)
  - We implement OLS/Logit/Probit
  - In this presentation, focus only on Logit
- Simulation to compare the proposed method to:
  - Unweighted model
  - Weighted model with weights generated by iterative proportional fitting (IPF) raking, using command *ipfraking.ado presented in a 2014 Stata Journal article by Kolenikov, S.*

The Method

- Imbens and Lancaster (1994) “Combining Micro and Macro Data in Microeconometric Models” *Review of Economic Studies*
- Hellerstein and Imbens (**hereafter H&I**) (1999) “Moment Restrictions From Auxiliary Data by Weighting” *Review of Economics and Statistics*
- H&I use moment restrictions from auxiliary/population data to (implicitly) re-weight survey data
  - Generally requires uncentered first, second and cross moments

The Method

- Auxiliary data are assumed to represent the population
- H&I differs from conventional (e.g., raking) methods because
  - Conventional methods create general purpose weights
  - H&I simultaneously estimates coefficients of the model of interest, and generates model-specific weights by matching sample moments to population moments
  - However, H&I can also be used in generating general purpose weights when model is not specified
- Can be extended to virtually any GMM model
- Suitable for auxiliary data on continuous and discrete variables

The Method

- Moment restrictions

$$E\ \rho(y,x,\beta,\lambda) = E\begin{bmatrix}\rho_1(y,x,\beta,\lambda)\\ \rho_2(y,x,\lambda)\end{bmatrix} = E\begin{bmatrix}\frac{x\ f(\beta' x)}{1+e^{\lambda' h(y,\ x)}}\\ \frac{h_m(y,\ x)}{1+e^{\lambda' h(y,\ x)}}\end{bmatrix} = 0$$

  ρ_1 : weighted score functions from log-likelihood for Logit/Probit model (or weighted normal equations for OLS)

  ρ_2 : weighted distance of individual observation for each weighting variable from the respective auxiliary/population moment

- GMM chooses -**simultaneously**-

  β**:** the coefficients of the model of interest to minimize its weighted criterion function

AND

  λ **:** to make the weighted moments in the sample as close to that (those) in the population as possible

The Method

- In: $E\left[\begin{matrix}\frac{x\ f(\beta' x)}{1+e^{\lambda' h(y,\ x)}} \\ \frac{h_m(y,\ x)}{1+e^{\lambda' h(y,\ x)}}\end{matrix}\right] = 0$

  $h_m(y,\ x)$: deviates the survey variables from their respective auxiliary/population moments

  - $E[weighted\ h_m(y,\ x)] = 0$

- **For example,** age, female, and age*female are weighting variables
- then,

  $h_{Age} = Age_i - \overline{Age_{pop}};\ i = 1....n$

  $h_{female} = female_i - \overline{female_{pop}}\ ;\ i = 1....n$

  and

  $h_{female*age} = female * age_i - \overline{female * age_{pop}}\ ;\ i = 1....n$

Overview of Results

- H&I improves precision of estimates in small simple random samples
- With sufficient auxiliary data on $\mathbf{x}$ and y, H&I performs better in a setting with biased sampling based on observables than unweighted regression and ipfraking
  - Largely comes from variance reduction and sometimes moderately from bias reduction
- With insufficient auxiliary data on $\mathbf{x}$ and particularly, no auxiliary data on y leads the unweighted method performing better than H&I and ipfraking

Simulation Strategy

- Computer generated population
  - Logistic distribution
  - **Five regressors:** three are continuous ($\mathbf{x}$) and two are discrete ($\mathbf{d}$)
  - Size: $100,050$
  - Correlation among regressors:

Variables Name	x_1	x_2	x_3	d_1	d_2
x_1	1				
x_2	0.50	1			
x_3	0	0.20	1		
d_1	0.48	0.32	0.28	1	
d_2	0	0.43	0	0	1

Simulation Strategy

- Computer generated population
  - y is generated with $pr(y = 1) = 0.5$
  - **The Logit probability:** $pr(y = 1) = \frac{exp(\beta' x)}{1+exp(\beta' x)}$
  - $\beta' x = 2 + 5x_1 - 2x_2 + 3x_3 + d_1 - 7d_2$
- Sample
  - $n = 200$, 500 and 2500
  - Selection on $x(s)$ and/or y variables (selection on y sometime called "choice based sampling" or "endogenous sampling") and simple random sample
  - Selection on y can happen in a sample of 500 observations if we draw 300 observations with $y = 0$ and 200 observations with $y = 1$
- **Iterations:** 1000

Simulation Strategy

- Various models based on auxiliary data used

<table>
<tr><th rowspan="2">Models</th><th colspan="2">Moments used</th><th rowspan="2">comments</th></tr>
<tr><th>x & y</th><th>Only x</th></tr>
<tr><td>1</td><td>First</td><td></td><td rowspan="4">Theoretical world, assuming moments are available from auxiliary data on each variable</td></tr>
<tr><td>2</td><td>First & Second</td><td></td></tr>
<tr><td>3</td><td>First, Second & Cross</td><td></td></tr>
<tr><td>4</td><td>First & Cross</td><td></td></tr>
<tr><td>5</td><td>Same as model-3 but no moments on x_2 & d_2</td><td></td><td rowspan="5">More practical world, assuming moments may not be available on each variable</td></tr>
<tr><td>6</td><td>Same as model-4 but no moments on x_2 & d_2</td><td></td></tr>
<tr><td>7</td><td></td><td>First, Second & Cross</td></tr>
<tr><td>8</td><td></td><td>First & Cross</td></tr>
<tr><td>9</td><td>Same as model-3 but without first moment on d_2, cross moment on $x_1 d_1$ and second moment on x_3</td><td></td></tr>
</table>

Simulation Results (sample only on y)

- Results from 200 observations are not presented
- Size 500 & 2500
- 60% observations with $y = 0$ and 40% with $y = 1$
- All statistics presented under various models unless otherwise mentioned are **ratios of mean squared errors (MSE)**
- If any ratio is less than one, e.g., 0.31, it means the MSE of the method in the numerator is only 31% of that of the denominator
- In subsequent tables, *if any ratio is marked as 'Black," it is better compared to the benchmark method, & if it is "Red," it is worse*

Simulation Results (sample only on y)

- **Model**-3 (first, second and cross moments of **x**, **d** and y)

	H&I/Unweighted	ipf/Unweighted	H&I/ipf	H&I/Unweighted	ipf/Unweighted	H&I/ipf
		n=500			n=2500	
const	0.31	0.48	0.65	0.06	0.10	0.64
x_1	0.81	1.08	0.75	0.72	0.92	0.78
x_2	0.69	1.00	0.69	0.63	0.92	0.69
x_3	0.63	0.93	0.68	0.51	0.83	0.62
d_1	0.41	0.57	0.72	0.39	0.52	0.75
d_2	0.69	0.91	0.76	0.60	0.73	0.82

- **Model**-5 (first, second and cross moments of all but x_2, d_2)

	H&I/Unweighted	ipf/Unweighted	H&I/ipf	H&I/Unweighted	ipf/Unweighted	H&I/ipf
		n=500			n=2500	
const	0.59	0.60	0.99	0.12	0.13	0.96
x_1	1.02	1.05	0.97	0.97	1.01	0.96
x_2	1.05	1.04	1.01	1.02	1.03	1.00
x_3	0.93	0.96	0.96	0.82	0.91	0.90
d_1	0.79	0.83	0.95	0.81	0.82	0.99
d_2	1.07	1.06	1.01	1.04	1.04	1.00

Simulation Results (sample only on y)

- **Model**-9 (Have some moments on each variable, but not all of them)

	H&I/Unweighted	ipf/Unweighted	H&I/ipf	H&I/Unweighted	ipf/Unweighted	H&I/ipf
		n=500			n=2500	
const	0.59	0.50	1.19	0.12	0.10	1.17
x_1	0.93	1.10	0.84	0.78	0.93	0.84
x_2	0.76	1.02	0.75	0.67	0.91	0.73
x_3	0.82	0.95	0.87	0.62	0.83	0.75
d_1	0.55	0.57	0.97	0.56	0.58	0.97
d_2	0.83	0.88	0.94	0.76	0.74	1.02

Simulation Results (Sample on x_3 and d_1)

- Size 500 & 2500
- Four strata based on values of x_3, and d_1

	$d_1 = 0$	$d_1 = 1$
$x_3 < \bar{x_3}$	Oversampling (by 56%)	Under-sampling (by 36%)
$x_3 \geq \bar{x_3}$	Under-sampling (by 36%)	Oversampling (by 56%)

Simulation Results (Sample on x_3 and d_1)

- **Model**-3 (first, second and cross moments of **x**, **d** and y)

	H&I/Unweighted	ipf/Unweighted	H&I/ipf	H&I/Unweighted	ipf/Unweighted	H&I/ipf
		n=500			n=2500	
const	0.56	0.89	0.63	0.45	0.68	0.66
x_1	0.86	1.18	0.72	0.75	1.06	0.71
x_2	0.73	1.12	0.65	0.68	1.10	0.62
x_3	0.70	0.99	0.71	0.64	0.91	0.71
d_1	0.42	0.61	0.68	0.43	0.65	0.66
d_2	0.69	0.96	0.73	0.62	0.82	0.75

- **Model**-5 (first, second and cross moments of all but x_2, d_2)

	H&I/Unweighted	ipf/Unweighted	H&I/ipf	H&I/Unweighted	ipf/Unweighted	H&I/ipf
		n=500			n=2500	
const	1.00	1.05	0.95	0.91	0.98	0.93
x_1	1.06	1.14	0.93	1.02	1.13	0.90
x_2	1.08	1.16	0.93	1.06	1.16	0.91
x_3	0.91	0.99	0.91	0.86	0.97	0.88
d_1	0.88	0.97	0.92	0.90	1.00	0.90
d_2	1.11	1.15	0.96	1.10	1.17	0.94

Simulation Results (Sample on x_3 and d_1)

- **Model**-9 (Have some moments on each variable, but but not all of them)

	H&I/Unweighted	ipf/Unweighted	H&I/ipf	H&I/Unweighted	ipf/Unweighted	H&I/ipf
		n=500			n=2500	
const	1.06	0.92	1.15	0.92	0.70	1.32
x_1	0.96	1.21	0.79	0.90	1.07	0.85
x_2	0.76	1.19	0.64	0.73	1.10	0.67
x_3	0.89	1.09	0.82	0.77	0.91	0.85
d_1	0.66	0.71	0.94	0.65	0.69	0.93
d_2	0.95	1.02	0.93	0.86	0.83	1.05

- In simple random samples, H&I also performs better
  - Improvement entirely comes from reduced variance

The Stata Program

- **Syntax**
  - Suggestions for program name or further options appreciated

svywt *depvar indepvar(s)* [*if*] [*in*], **wtvar**(*varlist*) **moments**(*numlist*)[*options*]

options	Description
* **wtvar**(*varlist*)	list of variables to be used in matching sample moments to population moments
* **moments**(*numlist*)	list of respective moment values for the **wtvar**
popsize(#)	the size of the population from which the sample is drawn
model(*name*)	the name of the model, which could be logit or probit, the default is OLS
owobst()	weight only or bootstrap option defined by *wtonly* or *boot*, the default is as is the **model**()
noextract	suppress weights from showing up both as variables and statistics
nocompare	suppress comparison of sample means to weighted means and population means
noconstant	suppress constant both from regression model and instruments
nolog	suppress log from GMM

* **wtvar**() and **moments**() are required.

Example with Real Data

- Stata's sample NHANES-II data
  - NHANES-II - no StatCan data because of RDC access limitations
- Toy/illustrative model
- Two different sets of *auxiliary data:*
  1. *Implied moments based on NHANES-II weights*
  2. *Moments (count totals) used by Kolenikov, S. in a 2014 Stata Journal article illustrating ipfraking from projected 2011 Census data*
- *2011 Census data is a completely different set of auxiliary data than from which NHANES-II data are sampled*

```
. svywt obesity i.age20_39 i.age40_59 i.female i.black ///
>i.age20_39#i.female i.age40_59#i.female i.black#i.female, ///
>wtvar(male_age20_39 male_age40_59 female_age20_39 female_age40_59 ///
>region1 region2 region3 female black orace) ///
>mom(0.2364 0.1601 0.2484 0.1754 ///
>0.2069 0.2489 0.2653 0.5206 0.0955 0.0253) ///
>mo(logit) pop(117157513)

Step 1
Iteration 0:   GMM criterion Q(b) =   .0330127
Iteration 1:   GMM criterion Q(b) =  .00062789
Iteration 2:   GMM criterion Q(b) =  5.982e-06
Iteration 3:   GMM criterion Q(b) =  1.482e-08
Iteration 4:   GMM criterion Q(b) =  2.936e-13

note: model is exactly identified
```

```
GMM estimation

Number of parameters =   18
Number of moments    =   18
Initial weight matrix: Unadjusted                 Number of obs   =     10,351

------------------------------------------------------------------------------------
                  |               Robust
                  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
obesity           |
       1.age20_39 |   -.328409   .1043887    -3.15   0.002    -.5330071   -.1238108
       1.age40_59 |     .17323   .1058513     1.64   0.102    -.0342347    .3806946
         1.female |    .482482   .0910516     5.30   0.000      .304024    .6609399
          1.black |   .2891128   .1455402     1.99   0.047     .0038593    .5743663
                  |
  age20_39#female |
             1 1  |  -.2916567   .1356623    -2.15   0.032      -.55755   -.0257635
                  |
  age40_59#female |
             1 1  |  -.1878141   .1371215    -1.37   0.171    -.4565674    .0809391
                  |
     black#female |
             1 1  |   .6830749   .1795611     3.80   0.000     .3311415    1.035008
                  |
            _cons |  -1.907744    .071204   -26.79   0.000    -2.047301   -1.768186
------------------+-----------------------------------------------------------------
xb                |
  W_male_age20_39 |  -1.632836     .05756   -28.37   0.000    -1.745652    -1.52002
  W_male_age40_59 |  -1.756258    .073151   -24.01   0.000    -1.899631   -1.612884
W_female_age20_39 |  -1.486732   .0537084   -27.68   0.000    -1.591998   -1.381465
W_female_age40_59 |  -1.656894   .0679773   -24.37   0.000    -1.790127    -1.52366
        W_region1 |   .1697596   .0673058     2.52   0.012     .0378427    .3016766
        W_region2 |   .3775347   .0606526     6.22   0.000     .2586578    .4964116
        W_region3 |   .2543315   .0619542     4.11   0.000     .1329036    .3757595
         W_female |  -.0635953   .0414605    -1.53   0.125    -.1448564    .0176659
          W_black |   .2217372   .0661167     3.35   0.001     .0921509    .3513235
          W_orace |  -.3378578   .1980346    -1.71   0.088    -.7259984    .0502828
------------------------------------------------------------------------------------
```

```
Instruments for equation first: 0b.age20_39 1.age20_39 0b.age40_59 1.age40_59
0b.female 1.female 0b.black 1.black 0b.age20_39#0b.female 0b.age20_39#1o.female
1o.age20_39#0b.female 1.age20_39#1.female 0b.age40_59#0b.female
0b.age40_59#1o.female 1o.age40_59#0b.female 1.age40_59#1.female
0b.black#0b.female 0b.black#1o.female 1o.black#0b.female
1.black#1.female _cons

Instruments for equation eqn1: _cons
Instruments for equation eqn2: _cons
Instruments for equation eqn3: _cons
Instruments for equation eqn4: _cons
Instruments for equation eqn5: _cons
Instruments for equation eqn6: _cons
Instruments for equation eqn7: _cons
Instruments for equation eqn8: _cons
Instruments for equation eqn9: _cons
Instruments for equation eqn10: _cons
```

```
                           hiweight
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .1642301       .1480108
 5%     .1782058       .1480108
10%     .1877105       .1480108       Obs              10,351
25%     .2214596       .1480108       Sum of Wgt.      10,351

50%     .5260476                      Mean           .4341402
                        Largest       Std. Dev.      .1772344
75%     .5781396       .7197252
90%     .6181775       .7197252       Variance        .031412
95%     .6386414       .7197252       Skewness      -.4250065
99%     .6763764       .7197252       Kurtosis       1.402055

                          normweight
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .3782881       .3409287
 5%     .4104798       .3409287
10%      .432373       .3409287       Obs              10,351
25%     .5101108       .3409287       Sum of Wgt.      10,351

50%       1.2117                      Mean                  1
                        Largest       Std. Dev.      .4082423
75%     1.331689       1.657817
90%     1.423912       1.657817       Variance       .1666618
95%     1.471049       1.657817       Skewness      -.4250065
99%     1.557968       1.657817       Kurtosis       1.402055

                           popweight
-------------------------------------------------------------
      Percentiles      Smallest
 1%     4281.643       3858.792
 5%     4646.004       3858.792
10%     4893.802       3858.792       Obs              10,351
25%     5773.675       3858.792       Sum of Wgt.      10,351

50%     13714.59                      Mean           11318.47
                        Largest       Std. Dev.       4620.68
75%     15072.68       18763.96
90%     16116.51       18763.96       Variance       2.14e+07
95%     16650.03       18763.96       Skewness      -.4250065
99%     17633.81       18763.96       Kurtosis       1.402055
```

```
Qc[10,3]
                  Samp_avg  Weighted_avg   pop_moments
male_age2~39     .18220462         .2364         .2364
male_age4~59     .11709014         .1601         .1601
female_ag~39     .19862815         .2484         .2484
female_ag~59     .13051879         .1754         .1754
     region1     .20249251         .2069         .2069
     region2     .26799343         .2489         .2489
     region3     .27562554         .2653         .2653
      female     .52516665         .5206         .5206
       black      .1049174         .0955         .0955
       orace      .0193218         .0253         .0253
r; t=224.63 18:39:42
```

- Comparison of estimates using NHANES-II (implicit) auxiliary data

	Unweighted		NHANES weighted		NHANES weighted ipf		NHANES weighted H&I	
	Coef	S.E.	Coef	S.E.	Coef	S.E.	Coef	S.E.
age20_39	-0.34	0.10	-0.35	0.12	-0.35	0.12	-0.33	0.10
age40_59	0.17	0.11	0.15	0.12	0.15	0.12	0.17	0.11
female	0.49	0.09	0.46	0.10	0.46	0.10	0.48	0.09
black	0.26	0.13	0.29	0.16	0.29	0.16	0.29	0.15
age20_39#female	-0.28	0.13	-0.31	0.15	-0.31	0.15	-0.29	0.14
age40_59#female	-0.19	0.14	-0.20	0.15	-0.20	0.15	-0.19	0.14
black#female	0.67	0.17	0.66	0.20	0.66	0.20	0.68	0.18
cons	-1.90	0.07	-1.91	0.08	-1.91	0.08	-1.91	0.07

- Comparison using 2011 census moments same as Kolenikov, S. (2014)

	Unweighted		NHANES weighs		2011 moments ipf		2011 moments H&I	
	Coef	S.E.	Coef	S.E.	Coef	S.E.	Coef	S.E.
age20_39	-0.34	0.10	-0.35	0.12	-0.26	0.13	-0.11	0.12
age40_59	0.17	0.11	0.15	0.12	0.22	0.13	0.38	0.12
female	0.49	0.09	0.46	0.10	0.49	0.11	0.57	0.11
black	0.26	0.13	0.29	0.16	0.38	0.17	0.32	0.15
age20_39#female	-0.28	0.13	-0.31	0.15	-0.41	0.17	-0.43	0.16
age40_59#female	-0.19	0.14	-0.20	0.15	-0.23	0.17	-0.31	0.15
black#female	0.67	0.17	0.66	0.20	0.65	0.21	0.71	0.19
cons	-1.90	0.07	-1.91	0.08	-1.99	0.08	-2.10	0.08

- Comparison of weights from different auxiliary data

	Mean	Std. Dev.	Min	Max
	NHANES (implicit) moments			
NHANES	11318	7304	2000	79634
ipfraking	11318	7305	2000	79806
H&I	11318	4623	3693	18800
	2011 Census moments			
ipfraking	22055	19227	4050	338675
H&I	22055	17561	5679	100453

Conclusions

- H&I performs very well if appropriate moments are provided in the restrictions
- Can perform worse than unweighted without appropriate moment restrictions, which is also true of *ipfraking (or weighting in general)*
- *The command we develop is very easy to use*

Thanks You

References

Hellerstein and Imbens (1999). "Moment Restrictions From Auxiliary Data by Weighting" *Review of Economics and Statistics.*

Imbens and Lancaster (1994). "Combining Micro and Macro Data in Microeconometric Models" *Review of Economic Studies.*

Kolenikov, S. (2014). "Calibrating Survey Data using Iterative Proportional Fitting (raking)" *The Stata Journal, 14(1), 22-59.*