Applied Econometrics for Practitioners:
Using Household Surveys to Inform Policy in Nepal

Dean Jolliffe, Hiroki Uematsu, Ganesh Thapa
Kathmandu, Nepal
01/22/2018

1

Announcements / reminders

1. Assignment due on Jan 26
2. Next TA class on January 23

We will start today where we left off last Friday, on the OLS assumptions, 1-5. We stopped at:

$$\text{Weight Loss} = \alpha + \beta \text{WeightWatchers} + \varepsilon$$

What's in ε
…that's not correlated with WeightWatchers?
…that is correlated with WW?
Nepal Example?

2

Outline for Sample Design issues

- Sample Frame & **Simple Random Sample**
- Describe **Stratification**, purpose 1: reduce variance, algebraic result & graphical illustration.
- Stratification, purpose 2: ensure sample coverage. Eg. of effect on **weights & means**.
- Describe **Multi-stage** design, purpose & derive result of increased variance. Examples showing effect and how to program.

3

Sample Frame

- Defines the population under consideration
- Constrains Inferences

Examples:

- Census data
- List of registered voters
- Phone book
- Land registry, housing structure registry
- Existing survey sample (follow up subsample)

4

Simple Random Sample - SRS

- All observations from the frame have equal probability of selection
- National lottery, selection of random digits
- Workers born on the 5th of the month in a labor force survey. (Any problems/concerns?)
- What is the relationship between the sample and the population?

5

Advantages of SRS

- Conceptually straightforward.
- Self-weighted samples (1/p is constant for all i)
- Inverse prob. of selection is expansion factor
- Sample mean = estimate of population mean
- Statistics derived from basic Stats books

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(y_i) \quad \text{and} \quad \bar{y} = \bar{Y}$$

$$\mathrm{Var}(\bar{y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

6

Disadvantages of SRS

- Can't guarantee coverage of particular subpopulations. Comparison of regions (see stratification)
- Low-probability outcomes hard to measure (see stratification)
- Expensive (at least for face-to-face interviews, see multi-stage)

7

Stratification - explanation

- Basic idea: Prior to any randomization, partition the sample frame into 'useful' sub-units (eg. 7 provinces, previously agri-eco regions/belts).
- Examples: Geographic regions, race, gender, income proxy, schooling
- The strata need to fully cover the frame (or the reference population is redefined)
- A fixed number of observations are then randomly drawn from within each stratum.
- *Conceptually equivalent to drawing multiple, independent random samples. (1 for each stratum.)

8

Stratification - purpose

- What's the purpose of stratifying a frame?
- Possible to improve precision (Textbook reason, we'll look at an algebraic expression for this result. We won't derive it.)
- Desire to estimate characteristic of a small sub-population or somewhat infrequent event (Domain of analysis, perhaps the most important reason)
- Administrative convenience - survey preparation may need to vary over regions

9

Stratification - Ensuring sufficient sample sizes of subpopulations

- Assume that you want to examine the differences between people who had "very low food security" (VLFS, 4%, 2006) and the rest of the population.
- SRS => In a sample of 1000, we'd expect 40 to be VLFS. (Maybe more, maybe less) Too small.
- Stratify on some characteristic that we hope is correlated with hunger. Say the frame contains an estimate of income. Oversample low income.

10

Stratification - How does over sampling affect inference?

- Over sampling => no longer the case that all observations in the frame have an equal probability of being selected.
- In VLFS example, low-income people have a higher probability of being selected.
- Consider the following example: Sample frame contains estimates of income. Say 30% estimated to be low income.

11

Stratification - How does over sampling affect inference?

FRAME			
Income		**Stratum 1**	**Stratum 2**
10, 10, 10,		10, 10, 10,	3, 3, 3, (low)
9, 9, 9,	STRATIFY SAMPLE TO	9, 9, 9,	2, 2, 2, (low)
8, 8, 8,	ENSURE POOR	8, 8, 8,	1, 1, 1 (low)
7, 7, 7,	ARE REPRESENTED	7, 7, 7,	
6, 6, 6,		6, 6, 6,	
5, 5, 5,		5, 5, 5,	
4, 4, 4,		4, 4, 4,	
3, 3, 3, (low)			
2, 2, 2, (low)		P=3/21	P=3/9
1, 1, 1 (low)		W=7	W=3

Choose 3 obs from each stratum. (why?) Low income stratum has 3 in 9 chance, high income 3 in 21. Therefore, we need to give smaller weights to low-income draws so that weighted sample estimates will reflect characteristics of the population.

12

Stratification - How does over sampling affect inference?

- Say our stratified random draw gives (1,1,3,5,6,9).
- Sample avg=25/6 (4.16)
- What does the sample average tell us about the population average?
- Nothing. (Assuming Prob. of selection not equal)
- Weighted sample average: $\sum \omega_i x_i / \sum \omega_i$
- (1*3 + 1*3 + 3*3 + 5*7 + 6*7 + 9*7) / (30) = 5.16
- Over sampling => differing weights. Weights allow us to draw inferences about POP from the sample.
- Note: If weights the same for all, then sample is called self-weighted

13

Sample weight example (We'll look at stratification example later)

NLSS-2011

Nepal Living Standard Survey

http://cbs.gov.np/nada/index.php/catalog/37

weights.do Help File

14

Comment (Stata weights example)

- Variance and standard deviation of the distribution of y

$$\text{Var}(y) = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2, \text{ Std. Dev}(y) = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

- Variance and standard error of the mean of y, Assuming SRS

$$\text{Var}(\bar{y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2, \text{ SE}(\bar{y}) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

15

Stratification - an algebraic statement regarding precision.

Let $\hat{\sigma}_h^2 = \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (y_{i,h} - \hat{y}_h)^2$, or the sample variance of y in stratum h. Discuss the terms.

16

Stratification - precision (cont.)

Because each sample is an independent, random draw within each stratum, the variance of $\bar{y}$ for the full sample can be expressed as ...

$$\hat{\sigma}_{SS}^2 = \frac{1}{n} \sum_{h=1}^{H} \omega_h \hat{\sigma}_h^2 ,$$

where SS abreviates 'stratified sample', n is the sample size, H is the number of strata, and ω_h is the proportion of the sample in stratum h (n_h / n).

Or, in words, the sample variance of $\bar{y}$ is the weighted average of the variance in each stratum.

17

Stratification - precision (cont.)

With some algebraic manipulation of terms, $\hat{\sigma}_{SS}^2$ can be written in terms of $\hat{\sigma}_{SRS}^2$ (i.e. the variance of $\bar{y}$ if y_i had been drawn following a Simple Random Sample).

$$\hat{\sigma}_{SS}^2 = \hat{\sigma}_{SRS}^2 - \frac{1}{n} \sum_{h=1}^{H} \omega_h (\bar{y}_h - \bar{y})^2$$

Discuss terms, note that result => $\hat{\sigma}_{SS}^2 < \hat{\sigma}_{SRS}^2$.

Or, in words, stratifying the sample can improve the precision (reduce the variance) of the estimator.

18

Stratification - implications (cont.)

$$\hat{\sigma}^2_{SS} = \hat{\sigma}^2_{SRS} - \frac{1}{n}\sum_{h=1}^{H} \omega_h (\bar{y}_h - \bar{y})^2$$

=> Precision is increased as ....

... heterogeneity across strata ↑.

Or in other words the more $\bar{y}_h$ differs from $\bar{y}$.

... homogeneity within strata ↑.

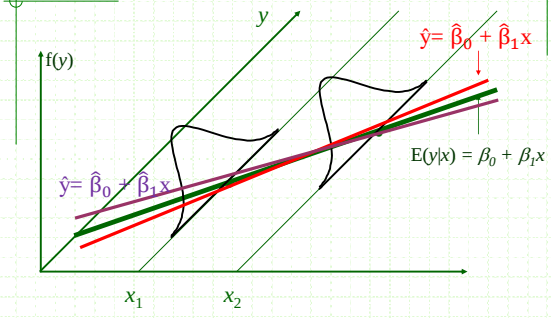
Decreased variance within each stratum ($\hat{\sigma}^2_h$) reduces the sum of the strata variances, or reduces $\hat{\sigma}^2_{SS}$.

Caveats: ω_h and relative magnitudes in practice.

19

Comment: N or n?

How precisely parameters are estimated is determined by sample not population size



20

Stratification - Illustration of increased precision

Overhead Slides to provide some intuition

$$\hat{\sigma}^2_{SS} < \hat{\sigma}^2_{SRS}$$

21