SUBSAMPLING STRATEGIES IN LONGITUDINAL SURVEYS

Jun Liu and Elvessa Aragon

Research Triangle Institute, Research Triangle Park, NC, 27709

Key Wrods: subsampling, longitudinal survey, unequal weighting effect, optimal sample allocation, Kuhn-Tucker conditions

1. INTRODUCTION

Subsampling in subsequent waves of data collection of a longitudinal study is usually used for multiple purposes. For example, a subsampling may be used for

- Minimizing the design effect and maximizing response rate,

or, it may be used for

- Minimizing field costs and maximizing response rate.

When conducting subsampling in a longitudinal study, a number of things should be considered. The primary concern is the possible reduction in the analytical capacity of the data as a result of subsampling. Such a reduction can be brought about in the following ways:

- Increased variation among analysis weights (subsampling will increase the weights of the subsampled groups),
- Possible reduction in the sample sizes of domains of interest, and
- Possible reduction in the overall weighted response rate (the cases removed through subsampling tend to have larger weights).

On the other hand, based on knowledge of the previous waves of data collection, one usually has the following information of the study design and sample:

- Inflated weights due to subsampling in previous waves,
- Inflated unequal weighting effect due to non-response adjustment in previous waves,
- Cohort members' response patterns in previous waves.

Features of longitudinal studies determines that subsampling in the follow-up waves of data collection usually is an act of balancing various considerations of the study goals and study realities. In this paper, we discuss one aspect of such follow-up studies: how to optimally subsample the study cohort. We first will introduce an example through it various points made above can be demonstrated. Then we will present our approach to the problem and the results.

2. AN EXAMPLE: NELS: 88/2000

The National Education Longitudinal Survey (NELS): 88/2000 is the 4th follow up of the survey. The original sample were students in the 8th grade in 1988 with refreshened samples in 1990 and 1992. There have been three previous follow-ups in 1990, 1992 and 1994. At the time of the 4th Follow-Up, several challenging issues emerged:

- There exists considerable variation in the weights due to weight adjustment and a subsampling that was carried out in the 3rd Follow-Up to increase the weighted response rate (Exhibit 1 provides for more details),
- The contact database has not been updated since the last follow-up in 1994 and the contact information for the non-respondents of the 3rd Follow-Up is 8 years old at the time of the 4th Follow-Up data collection,
- By the year 2000, many cohort members are in their mid-twenties and have moved away from where they were originally selected.

Given the mobility of the cohort and the age of the information in the contact database, tracing and locating was one of the most challenging aspects of the study. It was especially true for the the 3rd Follow-Up non/poor respondnets. Therefore, there was a desire to subsample out some of the 3rd Follow-Up poor/non-respondents to increase the weighted response rates. However, there were also analytical requirements that the 4th Follow-Up must satisfy. As a result, the subsampling problem became a problem of finding the best allocation of subsamples such that it would

- increase the locating and tracing success rate, therefore the weighted response rate,
- reduce the field cost for tracing and locating,
- yield sufficient number of interviews in key domains, and
- maintain the overall unequal weighting effect. as much as possible.

3. THE APPROACH

The goal is to achieve a balance among the competing priorities, using the available information. The strategy has the following steps:

- Stratify the follow-up sample by response propensities or interview cost and the size of the weights
- Allocate the sub-sample into strata such that the required response rates are met, key domain sizes are satisfied, and either the unequal weighting effect or the field cost is minimized
- Select the sub-sample with probability proportional to weight to further reduce weight variation among the sub-sample members

3.1 Decomposing the Unequal Weighting Effect

For a stratified sample, the unequal weighting effect for domain d can be decomposed as:

$$UWE_d = \frac{n_d \sum_h \sum_{i,j} (w_{hij} I_{dj})^2}{\left(\sum_h \sum_{i,j} (w_{hij} I_{dj})\right)^2}$$

$$= \sum_h \left(\frac{\hat{N}_{dh}}{\hat{N}_d}\right)^2 \left(\frac{n_d}{n_{dh}}\right) \left\{\frac{n_{dh} \sum_{i,j} (w_{hij} I_{dj})^2}{\left(\sum_{i,j} (w_{hij} I_{dj})\right)^2}\right\}$$

$$= \sum_h \left(\frac{\hat{N}_{dh}}{\hat{N}_d}\right)^2 \left(\frac{n_d}{n_{dh}}\right) UWE_{dh},$$

where,

UWE_{dh} is the unequal weighting effect for domain d and stratum h,

I_d is the domain indicator for domain d,

w_{hij} is the weight for sample member j in cluster i within stratum h,

$\hat{N}_d$ is the estimated population total for domain d, $\hat{N}_{dh}$ is the estimated population total for domain d within stratum h,

n_d is the sample size for domain d, and n_{dh} is the sample size for domain d within stratum h.

This formula was first derived by Folsom (1995) for the special case of two strata.

3.2 Sample Selection

Let w_{hij}^o be the original weight for sample member j in cluster i, stratum h. If the sample is selected with probability proportional to weight, then the selection probability within stratum h is

$$\pi_{hij} = n_h p_{hij} = \frac{n_h w_{hij}^o}{\sum_{i,j} w_{hij}^o} = \frac{n_h w_{hij}^o}{w_{h++}^o}.$$

Therefore, the new weight after the sub-sampling is

$$w_{hij}^n = w_{hij}^o \times \frac{w_{h++}^o}{n_h w_{h++}^o} = \frac{w_{h++}^o}{n_h} = \bar{w}_h^o,$$

a constant within a stratum! Since

$$UWE_{dh} = \frac{n_{dh} \sum_{i,j} (w_{hij}^n I_{dj})^2}{\left(\sum_{i,j} (w_{hij}^n I_{dj})\right)^2}$$

$$= 1 + CV_h^2(w_{hij}^n I_{dj}),$$

under the PPS to weight sampling, UWE_{dh} is 1. Thus,

$$UWE_d = \sum_h \left(\frac{N_{dh}}{\hat{N}_d}\right)^2 \left(\frac{n_d}{n_{dh}}\right) UWE_{dh}$$

$$= \sum_h \left(\frac{\hat{N}_{dh}}{\hat{N}_d}\right)^2 \left(\frac{n_d}{n_{dh}}\right).$$

Rewriting the above formula in terms of sub-sampling weights and sub-sampling rates, we have, for the overall unequal weighting effect,

$$UWE = \sum_h \left(\frac{w_{h++}}{w_{+++}}\right)^2 \left(\frac{1}{r_h^S}\right).$$

From the above formula, it is obvious that under the PPS proportional to weight sampling it is possible to use small sampling rates in small strata without inflating the overall unequal weighting effect drastically. It is also evident that in large strata small

sampling rates can have a significant inflation effect on the overall unequal weighting effect.

3.3 Response Rate Calculation

Denote the un-weighted response rate for domain d within stratum h as r_{dh}^{R} with

$$r_{dh}^{R} = \frac{\text{No of responders} \in \{\text{domain } d \wedge \text{stratum } h\}}{\text{All eligible sample members} \in \{\text{domain } d \wedge \text{stratum } h\}}.$$

The weighted response rate for domain d is

$$R_d = \sum_h \left(\frac{w_{dh++}}{w_{d+++}} \right) R_{dh},$$

with w_{dh++}^{n} and w_{d+++}^{n} being the sum of all weights in the respective domain and stratum.

Under the stratified PPS to weight sampling scheme, we also have

$$\begin{aligned} R_{dh} &= \frac{\sum_{(i,j)\in Response} w_{hij}^{n} I_d}{\sum_{all(i,j)} w_{hij}^{n} I_d} \\ &= \frac{\sum_{(i,j)\in Response} \bar{w}_h^{o} I_d}{\sum_{all(i,j)} \bar{w}_h^{o} I_d} \\ &= \frac{r_h^{R} n_{dh} \bar{w}_h^{o}}{n_{dh} \bar{w}_h^{o}} = r_{dh}^{R}, \end{aligned}$$

where n_{dh} is the sub-sample size for domain d within stratum h. Therefore,

$$R_{\text{overall}} = \sum_h \left(\frac{w_{h++}}{w_{+++}} \right) r_h^{R},$$

The advantage of expressing the target weighted response rate in terms of the unweighted response rates is that the unweighted response rates are easier to predict at the sample design stage.

3.4 Optimal Sample Allocation

If the objective is to achieve certain target weighted response rates at the same time minimizing the unequal weighting effect, the problem can be set up as a non-linear optimization over the space of the stratum sample sizes $\{n_h\}$.

Let the objective function be the unequal weighting effect given by

$$f(n_1, n_2 \ldots n_L) = \sum_h^{L} \left(\frac{w_{h++}}{w_{+++}} \right)^2 \left(\frac{n}{n_h} \right).$$

Let R_d^{C} be the target weighted response rates. The constraints on domain response rates are

$$g_d(n_1, n_2, \ldots n_L) = \sum_h \left(\frac{w_{dh++}}{w_{d+++}} \right) r_{dh}^{R} \le R_d^{C},$$

$$\text{for } d = 1,2,\ldots,D,$$

and on sample sizes are

$$l_h(n_1, n_2, \ldots, n_L) = n_h \ge 0, \quad \text{for } h = 1,2,\ldots,L,$$

Write

$$\begin{aligned} &L(n_1, n_2, \ldots, n_L, \lambda_1, \ldots, \lambda_D, \gamma_1, \ldots, \gamma_L) \\ &= f(n_1, \ldots, n_t) + \lambda_1 g_1(n_1, \ldots, n_t) + \ldots + \gamma_t l_t(n_1, \ldots, n_t). \end{aligned}$$

The Kuhn-Tucker conditions, which provide the necessary conditions for a local minimum, are given as

$$\frac{\partial L(n_1, n_2, \ldots, n_L, \lambda_1, \ldots, \lambda_D, \gamma_1, \ldots, \gamma_L)}{\partial n_h} = 0, \quad \text{for } h = 1,2,\ldots,L,$$

$$\frac{\partial L(n_1, n_2, \ldots, n_L, \lambda_1, \ldots, \lambda_D, \gamma_1, \ldots, \gamma_L)}{\partial \lambda_d} = 0, \quad \text{for } d = 1,2,\ldots,D,$$

$$\frac{\partial L(n_1, n_2, \ldots, n_L, \lambda_1, \ldots, \lambda_D, \gamma_1, \ldots, \gamma_L)}{\partial \gamma_h} = 0, \quad \text{for } h = 1,2,\ldots,L.$$

When the objective and constrains are convex functions, the Kuhn-Tucker conditions become sufficient as well (Chong, el at, 1996). The optimal solution should satisfy these conditions.

4. AN APPLICATION - NELS:88/2000

To implement the above approach, we first stratified the sample cohort by the 3rd Follow-Up subsampling strata (i.e., 2nd Follow-Up response status), the 3rd Follow-Up response status, the initial mailing status, and the availability of either the sample members' or their parents' Social Security Numbers

(SSNs). Exhibit 2 provides more details on the stratification. The purpose of this stratification is

- through the 2nd Follow-Up response status — to identify cases with large weights due to subsampling conducted in the 3rd Follow-Up,
- through the 3rd Follow-Up response status — to identify cases with older contact information and that would be less likely to respond,
- through initial mailing status and availability of SSN to identify cases for which tracing and locating would be difficult,

We then constructed the unequal weighting effects for each stratum defined above, for key analytical domains, and for the overall sample. We also constructed a cost function based on the estimated field cost of tracing, locating and non-response conversion. The constraints are setup for weighted response rates and domain sample sizes. We then minimized the unequal weighting effect and the cost function iteratively. Once an optimal solution is obtained, we selected a stratified sample with probability proportional to the weight. The cost and response rate assumptions are listed in Exhibit 3. The results of the subsampling selection is summarized in Exhibit 4. One point worth mentioning is that there is no appreciable difference between the overall unequal weighting effect in Exhibit 2 and 4 even though the subsampling removed 647 hard to interview subjects from the sample. The number of strata with large UWEs is also fewer in Exhibit 4.

5. DISCUSSIONS AND CONCLUSIONS

It can be argued that the overall unequal weighting effect reported in Exhibit 4 is an underestimate of the true unequal weighting effect, because the realized sampling rate in some of the strata are zero. However, since the individual stratum contribution to the overall unequal weighting effect is weighted by the proportion of the population in the stratum, in practice, the unequal weighting effect contribution of small strata will be smaller. More research in this area is needed and we are planning to conduct simulations to verify this conjunction. Another aspect of this approach that needs further research is the possible biases caused by selecting subsamples using proportional to weight method.

In summary, we have presented a formula that decompose the unequal weighting effect into contributions from individual strata and as a function of the stratum sample sizes. We have presented a procedure for obtaining the optimal allocation under competing priorities of the study. Through an example, it has been demonstrated that the procedure worked well. The procedure has utilities in other situations

6. REFERENCES

Chong, E.K.P. and Zak, S.H. (1996). *An introduction to Optimization*. Wiley, New York

Folsom, R.E. (1995). Private communication.

Exhibit 1. NELS Third Follow-Up Subsampling

3FU Subsampling Stratum	3FU Subsampling Rate	2FU Sample		3FU Sample	
		Sample Size	Mean Weight	Sample Size	Mean Weight
Excluded	0.00	731	184	0	0
Nonrespondents	0.15	288	197	43	1,319
Poor respondents	0.25	2,383	168	596	671
Dropouts	1.00	2,351	182	2,351	182
Inelig prior 92	0.90	212	214	191	238
Private school 88	0.80	2,984	108	2,387	135
Private school 90/92	0.80	122	376	98	469
Hispanic	0.90	1,629	118	1,466	131
API	1.00	874	76	874	76
Native American	1.00	132	163	132	163
Black high test	1.00	79	171	79	171
Black other	0.90	1,238	194	1,114	217
White low SES	1.00	1,295	157	1,295	157
White high SES	0.60	2,536	162	1,522	270
White mid SES	0.80	4,763	157	3,810	197
1FU freshened	0.30	4	93	1	370
2FU freshened	0.30	6	115	2	345
Other	0.40	8	159	3	424
Total		21,635	154	15,964	200

Exhibit 2. NELS:88/2000 Subsampling Stratification - Before Subsampling

2FU Response Status	3FU Response Status	Initial Mailing Status	SSN Status	Sample Size & Weight Statistics				
				Size	Mean	Range	Var	UWE
2FU Poor/Non-respondent	3FU Respondent	Received Response	w/ SSN	48	722	782	18091	1.04
			No SSN	7	916	849	109530	1.13
		No Response	w/ SSN	169	755	2486	63384	1.11
			No SSN	94	787	1824	96367	1.16
		Undelivered/Never sent	w/ SSN	60	776	1727	92102	1.15
			No SSN	75	852	3899	244382	1.34
	3FU Hostile Refusal	Undelivered/Never sent	w/ SSN	4	219	393	34721	1.72
			No SSN	16	167	296	7061	1.25
	3FU Non-respondent	No Response	w/ SSN	2	141	38	737	1.04
			No SSN	1	415	0	0	1.00
		Undelivered/Never sent	w/ SSN	12	327	763	56383	1.53
			No SSN	116	269	1677	65716	1.91
2FU Other	3FU Respondent	Received Response	w/ SSN	3919	199	3733	31611	1.80
			No SSN	52	200	798	23313	1.58
		No Response	w/ SSN	7900	193	6123	45393	2.22
			No SSN	207	189	2310	60850	2.71
		Undelivered/Never sent	w/ SSN	2296	206	4896	56680	2.33
			No SSN	74	271	2585	166342	3.27
	3FU Hostile Refusal	Undelivered/Never sent	w/ SSN	103	250	4687	260259	5.18
			No SSN	35	206	944	42806	2.01
	3FU Non-respondent	Received Response	w/ SSN	18	161	371	6963	1.27
			No SSN	2	162	135	9174	1.35
		No Response	w/ SSN	188	249	4345	221118	4.56
			No SSN	91	186	918	35108	2.02
		Undelivered/Never sent	w/ SSN	253	190	1964	50418	2.40
			No SSN	142	216	3560	115175	3.47
Overall				15884	215	6127	59480	2.29

Exhibit 3. NELS:88/2000 Subsampling Assumptions and Allocation

2FU Response Status	3FU Response Status	Initial Mailing Status	SSN Status	Stratum Size	Assumed Response Rate	Assumed Cost Factor	Assumed Sampling Rate
2FU Poor/Non-respondent	3FU Respondent	Received Response	w/ SSN	48	85%	1.05	1.00
			No SSN	7	75%	1.30	1.00
		No Response	w/ SSN	169	85%	1.50	1.00
			No SSN	94	70%	1.75	1.00
		Undelivered/Never sent	w/ SSN	60	75%	2.00	1.00
			No SSN	75	70%	3.50	1.00
	3FU Hostile Refusal	Undelivered/Never sent	w/ SSN	4	15%	15.00	0.05
			No SSN	16	10%	30.00	0.05
	3FU Non-respondent	No Response	w/ SSN	2	50%	3.50	0.35
			No SSN	1	15%	7.00	0.20
		Undelivered/Never sent	w/ SSN	12	40%	3.50	0.35
			No SSN	116	10%	10.00	0.30

2FU Other	3FU Respondent	Received Response	w/ SSN	3919	97%	1.00	1.00
			No SSN	52	87%	1.20	1.00
		No Response	w/ SSN	7900	90%	1.20	1.00
			No SSN	207	80%	1.50	1.00
		Undelivered/Never sent	w/ SSN	2296	85%	1.75	1.00
			No SSN	74	75%	3.00	1.00
	3FU Hostile Refusal	Undelivered/Never sent	w/ SSN	103	20%	10.00	0.15
			No SSN	35	15%	15.00	0.05
	3FU Non-respondent	Received Response	w/ SSN	18	75%	1.74	1.00
			No SSN	2	50%	2.50	1.00
		No Response	w/ SSN	188	60%	3.25	0.60
			No SSN	91	20%	5.00	0.30
		Undelivered/Never sent	w/ SSN	253	45%	3.25	0.59
			No SSN	142	15%	8.00	0.30

Exhibit 4. NELS:88/2000 Subsampling Stratification - After Subsampling

2FU Response Status	3FU Response Status	Initial Mailing Status	SSN Status	Sample Size & Weight Statistics				
				Size	Mean	Range	Var	UWE
2FU Poor/Non-respondent	3FU Respondent	Received Response	w/ SSN	48	722	782	18091	1.04
			No SSN	7	916	849	109530	1.13
		No Response	w/ SSN	169	755	2486	63384	1.11
			No SSN	94	787	1824	96367	1.16
		Undelivered/Never sent	w/ SSN	60	776	1727	92102	1.15
			No SSN	75	852	3899	244382	1.34
	3FU Non-respondent	No Response	w/ SSN	1	160	0	0	1.00
		Undelivered/Never sent	w/ SSN	4	576	498	69849	1.21
			No SSN	34	514	1564	110743	1.42
2FU Other	3FU Respondent	Received Response	w/ SSN	3919	199	3733	31611	1.80
			No SSN	52	200	798	23313	1.58
		No Response	w/ SSN	7900	193	6123	45393	2.22
			No SSN	207	189	2310	60850	2.71
		Undelivered/Never sent	w/ SSN	2296	206	4896	56680	2.33
			No SSN	74	271	2585	166342	3.27
	3FU Hostile Refusal	Undelivered/Never sent	w/ SSN	14	766	4669	1608650	3.74
			No SSN	2	181	143	10172	1.31
	3FU Non-respondent	Received Response	w/ SSN	18	161	371	6963	1.27
			No SSN	2	162	135	9174	1.35
		No Response	w/ SSN	77	425	4312	485888	3.69
			No SSN	26	341	882	81395	1.70
		Undelivered/Never sent	w/ SSN	121	264	1949	91426	2.31
			No SSN	37	444	3487	358286	2.82
Overall				15237	218	6127	61371	2.29