A COMPARISON OF WEIGHTS DERIVED FROM DIFFERENT MODELS

Jiahe Qian, NORC
1136 S. Washington St., Falls Church, VA 22046

KEY WORDS: Complex design, Model-based Weighting, Goodness-of-weighting, Simulation

In complex sampling, the inclusion rules of sample cases are sometimes defined indirectly. Accordingly, the case weights have to be derived from statistical models instead of to be imputed by the inverse of inclusion probability (Kish, 1992). The School Effectiveness Study (SES), an independent part of the National Educational Longitudinal Study of 1988 (NELS:88), is a survey of this kind.

When several statistical models are available, they would yield different sets of weights. To assess the effects of weighting, we compare the distributions obtained by applying the different sets of weights 1) with an empirical population database and 2) with populations drawn from a superpopulation in simulation.

1. Complex Sample Design and Model-based Weighting

In many survey sampling, the characteristics of aggregations, like schools and hospitals, are of interest. Particularly, in a longitudinal survey related to the transition of cases from one level of aggregations to another, sample design will become complex.

The School Effectiveness Study at 1990 contains a sample of 247 10th grade schools in the 30 largest Metropolitan Statistical Areas (MSAs) in the U.S. The objective of SES is to compare the effectiveness of education at school-level across different categories, like public vs private, etc. Hence, the target schools were set for 8 strata: 4 types of schools (Public, Catholic, NAIS[1], and Other Private) at two levels of urbanicity (Urban, Suburban).

To utilize existed longitudinal records of NELS:88 is the feature of the SES design. All the 10th grade schools on the list, from which first stage sample of schools was selected, must enroll at least one student who was selected in the NELS:88 study. The tenth-grade students, in the second stage sampling, were obtained from the sampled schools through two mechanisms: 1) a subsample of the students who were selected in the NELS:88, and 2) an augmentation sample of additional students.

Since the 10th grade schools in SES were identified as the students selected in NELS:88 transferred into them, the probability for a 10th grade school to enroll at least one student who was selected in NELS:88 was not defined directly. Let us call P_1, the chance for an aggregation to enroll one or more selected cases, as the *aggregation probability*. (B.D. Spencer and W. Foran, 1991) P_1 can be derived from different statistical models.

Three models for weighting. In SES, three different models were developed to impute P_1. They are Spencer-Foran model, Steve Kaufman model, and Qian-Frankel model. (M.R. Frankel and J. Qian, 1995)

The Spencer-Foran model is based on the probability model of hypergeometric distribution and the basis for the Qian-Frankel model is the probability model of binomial distribution, while Kaufman model estimates school weights by averaging of the student NELS:88 weights in a 10th grade school. See the appendix.

2. A Comparison of School Weights

Different statistical models would yield several sets of weights for aggregations correspondingly. Then, choosing proper weights in applications becomes fundamental.

To address the question of goodness-of-weighting, the performance of the weights in estimation, we examine the closeness between the sample distributions obtained by applying the different sets of weights and a population distribution.

To control the variance of estimates due to extreme large weights, all three sets of SES school weights were trimmed. The final weights were also adjusted by the raking process, matching weights with the marginal distributions specified by six variables obtained from Quality Education Data's (QED) database. However, the assessment process used only unraked weights; thus, the weighting effects can avoid being confounded by the effects of raking process.

2.1. Empirical data approach: comparing the weighted data with an empirical population database.

In the empirical data approach, QED database served as the "True Population", which contains 4,628 schools with tenth grades in the largest 30 MSAs. This was the population of schools from which the SES schools were selected. The QED file contains a number of school-level characteristics for each school on its list. Certain of the school-level variables available from the QED file are qualitative (nominal and ordinal) while

[1] Independent private schools are members of the National Association of Independent Schools.

others were quantitative (interval and ratio).

In assessing goodness-of-weighting, we measured the agreement between weighted sample distributions and the "True Population" of QED; we also examined a number of school-level estimates by applying weights derived from three models.

For the sample of 247 SES schools, the corresponding QED information for the schools was used in conjunction with the school weight. This produced 3 separate estimates for each variable.

Several statistical tests were employed in the assessment. For categorical variables, χ^2 statistic was used to test the agreement between weighted sample distributions and population distributions. Moreover, F test was used to check the homogeneity of two weighted sample distributions. For continuous variables, mean square error (MSE) also measured the goodness-of-weighting.

The χ^2 measures are shown in Tables 1. Besides, an F test procedure at $\alpha=0.25$ was used to provide some indication of the degree to which the various weights produced results that were different from one another. Table 1 shows that the Spencer-Foran and Qian-Frankel weighting models provided closer agreement between the weighted sample distributions and population distributions.

In addition to examining the performance of the three weighting models for categorical variables, table 2 shows the bias and root mean squared error for five continuous variables that were available from the QED file. By comparisons, we found that the Spencer-Foran model produced the lowest bias for five variables, while Kaufman model had the largest.

2.2. Simulation approach: comparing the weighted data with the populations drawn from a superpopulation.

Monte Carlo simulation approach is useful to approximate the exact sampling distribution of the weighted estimates by drawing a large number of samples from a fixed population for a given design. (C. Särndal, B. Swensson and J. Wretman, 1992) In assessing the effects of weighting in complex sampling, however, the Monte Carlo method has to be adapted to a new formation, different from regular approaches in finite population. Because of substantial computation and lacking auxiliary data for weighting, it is almost impossible to impute a separate set of weights for each realized sample. Instead of drawing a series of samples from the finite population, we keep the relationship between the sample and population, and weights fixed, then generate a series of sets of finite population variable values from a superpopulation. From realized samples of finite population, we estimate the variation of weighted estimates.

Let N be the size of finite population and n sample size; Y_k be the variable of interest and M_k the known auxiliary value, where $k=1,\dots N$. The N population values of $(y_1, y_2, \dots, y_N)$ are a realization of superpopulation ξ. y_k forms a general linear model

$$y_k = M_k + \varepsilon_k;$$

and, $E_\xi(\varepsilon_k) = 0$ and $V_\xi(\varepsilon_k) = \sigma_k^2$. Specifically, in the simulation for SES, assume $\varepsilon_k \overset{iid}{\sim} N(0, \sigma_k^2)$ and $\sigma_k = c \cdot M_k$, which set the coefficient of variation (cv) as a constant.

In assessing the effects of weighting under given sample design, we consider the criteria of unbiasdness and MSE.

In the simulation for SES, θ, calculated by M_ks, was the target statistic of superpopulation ξ. Estimate $\hat{\theta}_{w_r}$ was calculated by y_ks under a weight vector w_r for the SES sample. Given design p(.), sample and weights, $E_\xi(\hat{\theta}_{w_r} - \theta)$ measured the bias caused by design and weighting; and $E_\xi[(\hat{\theta}_{w_r} - \theta)^2]$ measured the mean square error. They were estimated by $\hat{b} = \frac{1}{A}\sum_{a=1}^{A}(\hat{\theta}_{w_r a} - \theta)$ and $\widehat{mse} = \frac{1}{A}\sum_{a=1}^{A}(\hat{\theta}_{w_r a} - \theta)^2$, where $\hat{\theta}_{w_r a}$ is the estimate in a^{th} repetition of the simulation.

The results of simulation approach in table 3 are consistent with the results obtained from the empirical data approach.

Conclusions. In general, the weights derived from different statistical models can be assessed by these issues: 1) model consistency with design as a probability sampling, 2) aggregation inclusion probability proportional to the number of selected cases, 3) bias in estimation (for continuous variables), 4) goodness-of-fit (for categorical variables), 5) cost of collecting data for weighting.

Summary of the Comparisons

Property	Spencer-Foran Weight	Kaufman Weight	Qian-Frankel Weight
(1)	good	N/A	good
(2)	good	N/A	good
(3)	small	largest	middle
(4)	good	medium	good
(5)	high	low	low

The table above summarizes the comparisons of three statistical models for weighting in SES. The method developed by Spencer and Foran provides weights that possess design validity; this is also true for the Qian-Frankel model. In addition, the Spencer-Foran and Qian-Frankel models conform more closely to intuition concerning how weights should behave in that larger tenth grade schools with more NELS:88 students will have higher probabilities of selection. This is not what is observed with the Kaufman model.

Appendix

The Spencer-Foran model. With the Spencer-Foran model, let eighth grade schools be primary sampling units (PSUs), and the students enrolled in those schools the secondary sampling units (SSUs), stratified into L strata of sizes N_{jl} for school j. Furthermore, consider the PSUs to be selected with replacement with selection probabilities proportional to a measure of size, and consider SSUs to be simple random subsamples of size n_{jl} selected without replacement in PSUs, with $n_{jl} > 0$ unless $N_{jl} = 0$. Also, consider that each aggregation contains one or more SSUs, although some aggregations may contain no SSUs that were selected into the sample.

In SES, tenth grade schools are the aggregations of interest. A PSU *feeds* an aggregation if at least one SSU from the PSU belongs to the aggregation. Now define B_{jl} as the number of SSUs in PSU j in stratum l that belong to the aggregation, and $B_j = \Sigma_{l=1}^{L} B_{jl}$. The number of SSUs in the aggregation is $\Sigma_1^A B_j$, where A is the number of PSUs that feed the aggregation.

In NELS:88	In Aggregation YES	NO	Total
YES	b_{jl}	$n_{jl}-b_{jl}$	n_{jl}
NO	$B_{jl}-b_{jl}$	$N_{jl}-B_{jl}-n_{jl}+b_{jl}$	$N_{jl}-n_{jl}$
Total	B_{jl}	$N_{jl}-B_{jl}$	N_{jl}

Given that PSU_j was selected, the probability model of hypergeometric distribution can be employed in the computation of the conditional probability that the number of SSUs in stratum l in PSU_j belong to the aggregation is b_{jl}. Then, the conditional probability that $b_{jl}=0$ can be calculated:

$H(n_{jl}, B_{jl}, N_{jl}) = [(N_{jl} - B_{jl})!(N_{jl} - n_{jl})!] / [N_{jl}!(N_{jl} - B_{jl} - n_{jl})!]$ if $N_{jl}-B_{jl}-n_{jl} \geq 0$, 0 otherwise, where N_{jl} is the number of SSUs in stratum l in PSU_j, and n_{jl} is the number of the selected SSUs in stratum l in PSU_j, if PSU_j has been selected.

Given that PSU_j was in the sample, the conditional probability that none of the SSUs in PSU_j that were selected belong to the aggregation is

$$\Pi_{k=1}^{3} H(n_{jl}, B_{jl}, N_{jl}).$$

Therefore, the *aggregation probability* P_1 equals

$$1 - \Pi_{j=1}^{A} [1 - S_j + S_j \Pi_{k=1}^{3} H(n_{jl}, B_{jl}, N_{jl})],$$

where S_j is the probability that PSUj was selected into the NELS:88.

The Kaufman model. As a means to reduce the cost of collecting feeder pattern data, following is an alternative estimator of the population total of X of interest,

$$\hat{X} = \sum_{k=1}^{N} 1/S_k \sum_{i=1}^{S_k} W_{i_k}^{88} I(i_k \in J \mid k \in K) I(k \in K) X_k,$$

which was proposed by Steven Kaufman and Bruce Spencer separately. Let us define $W_{i_k}^{88}$ as the weight of the student case selected in NELS:88, which approximately equals the reciprocal of the inclusion probability; J represents the NELS:88 student sample; K represents tenth grade schools, with size S_k, on the frame of the second stage, which contain at least one student selected in NELS:88 and i_k represents the student in tenth grade school k. In addition, we define $I(k \in K) = 1$ as k is in SES frame, 0 otherwise; and $I(i_k \in J \mid k \in K) = 1$ as i_k is selected in NELS:88 given k in the frame, 0 otherwise.

The estimate $\hat{X}$ will be unbiased if we can average over all possible NELS:88 samples and all possible SES samples; however, this estimate would not be unbiased *for a given NELS:88 sample* collected.

The Qian-Frankel model. As an alternative to the Kaufman model, Qian-Frankel model attempts to preserve some design-based properties of the Spencer-Foran model, while eliminating the need to obtain the feeder pattern information required by Spencer-Foran model.

Given the basic sample design, we assume that students in each of the tenth grade high schools possess similar characteristics related to the NELS:88 sample design, such as school type, area region, urbanicity, etc. Specifically, we assume that the chance for students in the same tenth grade school to have been included in NELS:88 are close. Note that the homogeneity in the chance for students in a tenth grade school to be included in the NELS:88 study is different from the assumption that students have the same chance to move to a tenth grade school.

Let J be a school in NELS:88; $P_k = P\{i_k \in J\}$ represents the chance of a student, in SES school k, to be included in NELS:88. The SES school k is of size S_k. In application, P_k may be estimated by the inverse of the average of the weights of the students selected in NELS:88: $\hat{P}_k = n_k C / \sum_{i=1}^{n_k} W_{i_k}^{88}$, where $W_{i_k}^{88}$ is the

NELS:88 weight for case i in school k and C is a constant determined by the sample design.

Let 'a student shifting to a high school' and 'being selected into the NELS:88 sample' be independent. As such, ξ_k, the chance for SES school k having n_k students selected in NELS:88, will form a binomial distribution. Therefore, the chance for SES school k having no students selected in NELS:88 can be estimated by , and also the probability that the SES school falls in the sampling frame is

$$\hat{P}(\xi_k > 0) = 1 - (1 - \hat{P}_k)^{S_k}.$$

As a result, the estimated weight for school k will be the reciprocal of .

References

Hammersley, J.M. and Handscomb, D.C. 1964, *Monte Carlo Methods*, London, Methuen.

Frankel, M.R. and Qian, J. 1995, "Sample Weighting for Multi-level Analysis in Panel Designs." Presented at 1995 meetings of American Educational Research Association. Kalos, M.H. and Whitlock, P.A. 1986, *Monte Carlo Methods*, New York, Wiley

Kish, L. 1965, *Survey Sampling*, New York, Wiley

Kish, L. 1992. "Weighting for Unequal P_i" *Journal of Official Statistics*, Vol 8, No 2, 183-200.

Särndal, C., Swensson, B., and Wretman, J. 1992. *Model Assisted Survey Sampling*, Springer-Verlag New York.

Spencer, B.D., and Foran, W. 1991. "Sampling Probabilities for Aggregations, with Application to NELS:88 and Other Educational Longitudinal Surveys," *Journal of Education Statistics* 16, 21-34.

Table 1. The χ^2 test for some variables in the QED file

	d.f.	Spencer-Foran Weight	Kaufman Weight	Qian-Frankel Weight
School type	2	$60.23^{K,Q}$	2091.53	328.14^{K}
Grade level	7	$131.26^{K,Q}$	524.87	298.03^{K}
Instruction dollars per pupil	6	78.86^{K}	277.17	$32.25^{K,S}$
MSA	29	282.57	$82.46^{S,Q}$	154.36^{S}
Urbanicity	1	161.87^{K}	1452.05	284.81^{K}
Personnel gender	1	$8.94^{K,Q}$	1631.49	85.54^{K}
# of students code	8	89.03^{K}	1159.55	130.09^{K}
# of teachers code	8	$52.73^{K,Q}$	1003.85	148.09^{K}
Enrollment change (building)	6	118.70^{K}	252.09	113.54^{K}
Region	3	62.57^{K}	515.37	127.00^{K}

Note: To compare the goodness-of-fit of the distribution of a certain variable between two sets of weights, an F-test, at $\alpha = .25$, is used to test the significance of difference. The letters S, Q, and K stand for Spencer-Foran, Qian-Frankel, and Kaufman weights, respectively. When a letter appears as a superscript in the cells, it means that the distribution under the weights of that column is fitted better than that under the weights represented by the letter in the superscription. For example, $94.86^{K,Q}$ in Table 1. shows that the distribution of teacher population code is fitted better under the Spencer-Foran weights than under the Kaufman or Qian-Frankel weights.

Table 2. The bias and root mean square errors
of the mean estimates for some variables in the QED file

Variable	Population Mean	Spencer-Foran W.		Kaufman Weight		Qian-Frankel W.	
		Bias	RMSA	BIAS	RMSA	BIAS	RMSA
# of White students	944.26	-54.42	115.39	190.33	199.57	64.47	117.60
# of Black students	273.94	-72.50	78.01	81.56	90.24	80.14	100.15
# of Hispanic students	190.52	-63.07	66.64	103.65	108.27	86.17	93.88
# of students	937.56	-6.21	79.79	696.31	699.82	352.14	368.52
# of teachers	51.59	1.82	4.85	29.51	29.69	13.51	14.56

Note: 1) The calculations for each statistic are based on those cases with nonmissing and nonzero values. 2) To include the impact of the complex sample design and weighting, the standard errors of the weighted means are calculated by the delta method.

Table 3. The bias and root mean square errors
of the mean estimates in simulation (repetition A=1000)

Variable	Population Mean	Spencer-Foran W.		Kaufman Weight		Qian-Frankel W.	
		Bias	RMSA	BIAS	RMSA	BIAS	RMSA
Coefficient of variation (CV)=0.2 in models:							
# of White students	944.26	-56.12	74.12	189.48	192.27	63.18	82.53
# of Black students	273.94	-72.92	73.26	81.36	82.29	79.77	80.90
# of Hisp. students	190.52	-63.14	63.22	103.67	103.99	86.15	87.07
# of students	937.56	-5.05	34.55	695.53	696.36	351.31	355.30
# of teachers	51.59	1.80	2.94	29.49	29.04	13.45	13.75
Coefficient of variation (CV)=0.8 in models:							
# of White students	944.26	-59.36	207.80	186.25	229.70	61.23	223.20
# of Black students	273.94	-72.34	77.96	82.45	96.56	81.39	102.80
# of Hisp. students	190.52	-63.20	64.73	103.19	108.48	85.73	100.55
# of students	937.56	0.98	144.23	692.68	706.66	350.64	413.69
# of teachers	51.59	1.62	9.75	29.16	30.00	13.03	17.27

Note: To include the impact of the complex sample design and weighting, an adjustment has been made to the standard errors of the weighted means by multiplying the standard errors by DEFT (L. Kish, 1965). DEFT were calculated by the delta method from QED data set.