EXAMINING THE ALTERNATIVE REGRESSION WEIGHTING FOR THE NATIONAL SURVEY OF COLLEGE GRADUATES

Elizabeth T. Huang, John M. Finamore, Patrick E. Flanagan and Thomas F. Moore
Elizabeth T. Huang, U.S. Bureau of the Census, Washington, D.C. 20233

Key Words: Regression weighting, Ratio adjustment, Relative efficiency

I. Introduction

The National Survey of College Graduates (NSCG) survey design is both cross-sectional and longitudinal in nature. As a cross-sectional study, the NSCG survey provides estimates of the size and characteristics of the science and engineering population of the nation for a point in time. The 1995 NSCG used the week of April 15, 1995 as its reference period. Longitudinally, the survey follows scientists and engineers, identified in the 1993 NSCG, to update estimates made from the 1993 survey and to examine changes in the nation's scientist and engineer's workforce throughout the decade. The follow-up surveys were conducted in 1995, 1997, and 1999.

The National Science Foundation has sponsored this survey of the college-educated population every decade since the 1960s. Prior to the 1993 survey, it was called the National Survey of Natural and Social Scientists and Engineers. Its sample design has been reevaluated and modified several times since its inception.

This research is motivated by Steinberg's (1992) recommendations on reweighting the NSCG using a regression approach (Moore (1997)). In his report on additional weighting methods for the 1989 National Survey of Natural and Social Scientists and Engineers, Joseph Steinberg recommended reweighting the series of surveys using poststratification with regression weighting. He was concerned with the "stratum 11" problem (the weight was extremely large for stratum 11 in comparison with the other 10 primary strata) and nonresponse bias.

In the NSCG design for the 1990s, we do not have the "stratum 11" problem. In addition, the weighted response rate in the 1993 NSCG was about ten percentage points higher than the response rate (70.6 percent) in the 1982 Survey of Scientists and Engineers (SSE). However, because of the longitudinal nature of the survey, the compound response rate is bound to decrease over time. Therefore the issue of nonresponse bias once again becomes a concern. Using regression estimation is one way for reducing bias associated with the nonresponse. (Fuller, Loughin and Baker (1994)).

It was proposed that we examine the benefits of applying Steinberg's recommendation to the NSCG surveys in 1995 and later. For both the 1993 and 1995 NSCG surveys, a ratio adjustment was used to control the NSCG sample back to the sampling frame. In this paper an alternative regression procedure is applied to the 1995 NSCG data. The estimated variances of the alternative regression estimates are then compared with the variances of the currently used ratio estimates.

II 1995 NSCG Sample Design and Estimation

I.A. Sample Design

NSCG is one of three surveys that are combined to create the Scientists and Engineers Statistical Data System (SESTAT). The other two surveys are the National Survey of Recent College Graduates (NSRCG) and the Survey of Doctorate Recipients (SDR). SESTAT has a target population of U.S. residents with a bachelor's degree or higher, who either have at least one degree in a science or engineering (S&E) field or are working in a science or engineering occupation, and as of the survey's reference period, are age 75 years or less.

The 1995 NSCG was selected from respondents (who have either an S&E degree or S&E occupation) to the 1993 NSCG, and respondents to the 1993 NSRCG.

The 1995 NSCG sample design is a stratified multiple phase unequal probability sample design. (Cox etc. (1997)).

The 1993 NSCG was selected from the 1990 Decennial Census Long Form sample respondents. Of the 148,932 complete interview cases, approximately 66,500 had an S&E bachelor's, master's, or foreign-earned PhD prior to April 1, 1990. A stratified multistage sample design was used for the 1993 NSRCG.

A total of 62,004 sample cases was selected for the 1995 NSCG. Due to cost considerations, a subsample of all mail nonrespondent cases was selected for computer assisted telephone interview (CATI) and computer assisted personal interview (CAPI). In total, there were 53,448 complete interview cases.

The strata for the 1995 NSCG were defined based upon demographic group, highest S&E degree, highest S&E major, and sex. The demographic group was a composite variable recording disability status,

---

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

citizenship, race/ethnicity, and institutionalization.

The overall weighted response rates varied by the demographic groups. The compound response rates (the product of the weighted response rates of 1993 NSCG and 1995 NSCG (1993 NSCG sample only)) for the demographic group are as follows (Finamore (1998)):

	(1) 1993 NSCG Response Rates	(2) 1995 NSCG Response Rates	(1)x(2) Compound Response Rates
Disabled Persons	78.9%	93.9%	74.0%
Hispanics	74.6%	93.7%	69.9%
Whites	81.7%	95.1%	77.7%
Blacks	70.9%	91.9%	65.1%
Asian & Pacific Islanders	81.5%	93.8%	76.5%
Native Americans	73.9%	91.2%	67.4%
Foreign Born U.S. Citizens	76.6%	92.2%	70.6%
Foreign Born Non-U.S. Citizens	64.9%	90.4%	58.7%
Total	79.8%	94.5%	75.4%

The overall weighted response rates of 1995 NSCG by gender are not much different; 94.4% for male and 93.8% for female. The weighted response rates by degree are 94.0% for Bachelors, 95.2% for Master, 92.0% for Professional and 93.9% for Doctorate. The weighted response rates by age are 91.4% for ages less than 30, 94.2% for ages between 30 and 59, and 95.7% for ages of 60 or more.

II.B. Estimation Procedure

Estimates in the 1995 NSCG are formed by inflating the responses of the interviewed persons in the NSCG to the national level. This is accomplished by assigning each sample person a final weight which is the product of the base weight (the product of the reciprocal of the probability of selection at multi-phase sample selections), subsampling adjustment factor, nonresponse adjustment factor and the ratio adjustment factor (Town (1996)). We defined the noninterview adjusted estimate in this paper as the estimate that uses the weight which is the product of the first three weighting components; and the ratio adjusted estimate as the estimate that uses the final weight.

The ratio adjustments were performed to control the 1995 NSCG sample back to the 1995 sampling frame. The pseudo strata were used as the adjustment cell for both the noninterview and ratio adjustments. The pseudo strata were formed from the sampling strata by further collapsing the small sampling strata. The ratio adjustment factor used in 1995 NSCG is calculated for each pseudo stratum. The control for each pseudo stratum used for the numerator of the ratio adjustment factor is estimated from 1995 NSCG frames. The denominator of the ratio adjustment factor is the weighted number (after noninterview adjustment weights) of the complete interview cases and the out of scope cases. VPLX (Variances from ComPLeX Surveys, Fay (1990)), a variance estimation software package (using replication methods) developed at the Census Bureau, was used for the variance estimation of the ratio adjusted estimates of the1995 NSCG survey variables.

A total of 160 replicates was created for the 1995 NSCG sample selected from the 1993 NSCG using the successive difference replication method (Fay & Train (1995)), and a total of 50 replicates was created (by WESTAT) for the 1995 NSCG sample selected from the 1993 NSRCG using the Jackknife (2 PSU per stratum) replication method.

III. Regression Weighting

In many sampling situations, the population means $(\bar{X}_1, \bar{X}_2, ..., \bar{X}_k)$ or totals of k auxiliary variables $(x_1, x_2, ..., x_k)$ are available, and the information can be used to improve the sample estimates. Common estimation procedures utilizing auxiliary information are ratio estimation, poststratification, regression estimation, and raking. The regression procedure is the most general procedure because it can incorporate all forms of the auxiliary variables (either discrete or continuous) in the regression estimator. Using regression estimation can reduce variance in sample estimates and has the potential for reducing bias associated with selective nonresponse. In this paper we applied an iterated regression procedure (Huang (1978), Huang & Fuller (1978)) to the 1995 NSCG data. The iterated regression procedure starts with the generalized regression estimator of the population mean of survey variable y in terms of the form $\sum_i w_i y_i$, where the initial regression weight w_i is a linear function of the auxiliary variables x_i's.

The initial regression weights may be negative. Such negative regression weights may then produce negative estimates of the population means known to be positive. The iterative regression procedure is designed to provide nonnegative regression weights w_i for a regression estimator of the form $\sum_i w_i y_i$.

The regression procedure is iterative, checking the weight w_i at each step against a user supplied criterion. The criterion, denoted by M ($0<M=<1$), is the maximum fraction by which any weight can deviate from the mean weight. If for any unit i whose regression weight does not satisfy the criterion, an adjusting factor g_i (a "bell" shaped function of distance (in a suitable metric)) is calculated for each observation which gives small factors to observations that have large or small regression weights, and a factor 1 if the observation is not far from

the population mean.That is, in the final regression estimator, the contribution of the regression estimation of the slope vector is reduced for observations that are far from the population mean of the auxiliary variables. The regression procedure will produce the regression weight w_i that is nonnegative for all sample units for suitable choice of M. The large sample distribution of the modified regression estimator was shown to be the same as that of the ordinary regression estimator under suitable conditions (Huang & Fuller (1978)). Once the regression weights are produced, they can be applied to any survey variable y in the form of $\sum_i w_i y_i$ to obtain the regression estimate of the population mean of the y's. For estimating the total Y, the regression weights for total, W_i, are the weights for the mean w_i multiplied by the population size N. ($W_i = w_i N$).

For a survey variable y, and k auxiliary variables of sample size n (x_{ij}, j=1,...,k, i=1,...n), the regression estimator of the total Y has the form $\sum_i W_i y_i$. The final regression weight W_i of sample unit i (i=1, . . ,n) for estimating the population total Y will have the following properties:

1. $W_i \geq 0$ for i = 1, 2, . . ., n;
2. $(1-M) \max_{1 \leq i \leq n} \{W_i\} \leq (1+M) \min_{1 \leq i \leq n} \{W_i\}$;
3. $\sum_{i=1}^{n} W_i = N$;
4. $\sum_{i=1}^{n} W_i x_{ij} = X_j$ j=1, ..., k.

where the parameter M, $0 < M \leq 1$, is specified by the user and is generally chosen in the interval [0.8,1.0]. X_j is the given population total for auxiliary variable x_j.

In addition to the array of observations and the population means (or totals) of the auxiliary variables, the sampling weights and the parameter M are also required as the input of the procedure.

The regression weighting procedure has been used for a number of large surveys by the Survey Group of the Iowa State University's Statistical Laboratory. (See Goebel (1976), Fuller, Loughin, and Baker (1994)).

The regression weighting procedure is applied to each demographic group using 1995 NSCG complete interview data for selected characteristics. For our study, we assumed that the frames of the 1995 NSCG are fixed, and the totals of the auxiliary variables for each demographic group were known. The parameter M, and the control totals X_j's are supplied for each demographic group. The auxiliary variables used in the regression estimate are similar to the auxiliary variables used in the current ratio adjusted estimate defined by the ratio adjustment cells. Beside the degree, gender, and S&E major field variables, age group variables were also used as auxiliary variables in the regression estimation.

The control totals are estimated from 1993 NSCG & 1993 NSRCG final weights. In our study we assumed that the control totals have no error. All the empirical work in this paper is conditional on the1995 NSCG frames which are the 1993 NSCG and 1993 NSRCG sample respondents that have an S&E degree or occupation and are U.S. residents under age 75.

The regression weights computer program (Huang (1983)) was modified to handle 20,000 observations with 50 independent and 50 dependent variables. The 1995 NSCG has 53,448 interviewed sample persons. The number of interviewed sample persons varied from 384 to 36,286 by the demographic groups.

For demographic group 3 (U.S.-born, nondisabled, noninstitutionalized White), the number of interviewed sample persons (36,286) is too large for the regression weights program. Therefore, we further split the interviewed sample persons in the demographic group White by Majors. The number of interviewed sample persons in demographic group 3 (White) by Majors varied from 2,143 to 11,992.

The regression estimates were computed for each demographic group and the demographic group White by Majors for the eight selected survey variables using the regression weights computer program.

The eight selected survey variables (y's) using 1995 NSCG data are indicator variables for employment status (work for pay, not looking for work, work closely related to the highest degree) and work activities (applied research, basic research, computer applications design & teaching, development, and professional services).

The auxiliary variables used for the regression of each demographic group (except the demographic group White) are indicator variables for gender, major, degree, and age group. A total of 14 auxiliary variables is used.

Since the demographic group White is further split by Majors, the auxiliary variables used in the demographic group White by Majors are gender, degree, and age group. A total of 10 auxiliary variables is used . For the demographic group White with no degree but Science & Engineering occupation, a total of eight auxiliary variables is used. These are indicators for gender, and age group.

The regression procedure was applied to each demographic group. The regression weights were obtained for each demographic group. It turns out that all regression weights are nonnegative for the initial regression.

IV. Evaluation Criterion

To compare the regression estimate with the current ratio estimate of the total for selected characteristics using the 1995 NSCG interviewed sample data, we compared their estimated variances by computing the ratio of the standard errors of the regression estimate and the ratio estimate assuming that the 1995 NSCG frames are fixed. In the 1980's, the estimation procedure for SSE did not use a ratio adjustment. The noninterview adjusted estimate was the final estimate in the 1980's SSE. We therefore compared the variances of the regression with that of the noninterview adjusted estimate and the current ratio estimate to measure the gain of the regression over the noninterview adjusted estimate and the current ratio estimate.

Originally, we planned to compute the estimated variance of the regression estimate using VPLX. However, the current version of VPLX does not have a procedure for computing the regression estimate and its estimated variance using replication methods. Alternative variance software was investigated. The WesVarPC (Brick, Broene, James, and Severynse (1997)) - a replication variance program by WESTAT, has a regression procedure except that there is no successive difference replication method.

The PC CARP software (Fuller, Kennedy, Schnell, Sullivan and Park (1966)) used a linearization method for the variance estimation in complex sample survey for a range of estimators. The PC CARP software was first used in computing the variance of the regression estimate. Later we found an equivalence of Fay's method with k=0.5 in WesVarPC and the successive difference replication method using VPLX when the current 160 replicates (created for the successive difference replication method) are used. This allowed us to use the current replicates and WesVarPC to estimate the variance of our regression estimates. There are 210 replicates for the 1995 NSCG sample cases. We used WesVarPC with the current replicates and the linear regression menu for each demographic group to compute the variance of the regression estimate. We used Fay's method (k=0.5) for the first 160 replicates, and the Jackknife method (JK2) for the last 50 replicates of all 1995 NSCG sample data. The full sample weight is the noninterview adjusted sample weight.

For each demographic group, the regression menu was used for each survey variable y on the auxiliary variables x's which are defined as the deviation from their population means. The estimate of the intercept from the WesVarPC regression run of each survey variable y is the regression estimate for the mean of y from each method. (Fay's method and JK2). The standard error of the regression estimate for the total y of each demographic group is the total number of the demographic group multiplied by the standard error of the intercept which is the square root of the sum of the variances of the intercept computed from the WesVarPC using two replication methods (Fay's method (k = 0.5) and Jackknife method).

In this study, all the variance estimates for different estimates are based on the same variance replication method.

V. The Comparison of the Regression Estimate with the Noninterview Adjusted Estimate

The relative efficiency of the regression estimate is computed as the ratio of the standard error of the regression estimate to the standard error of the noninterview adjusted estimate. The relative efficiency of the regression estimate is computed for different demographic groups (see table 1) and for the White group by Majors. In summary, the regression estimate is considerably better than the noninterview adjusted estimate for selected variables by demographic groups (or White demographic group by Majors). For the variable "working for pay," the relative efficiency ranges from 0.28 to 0.64 for different demographic groups, and 0.22 to 0.43 for the White demographic group by Majors. For the variable "not looking for work," the relative efficiency of the regression estimate ranges from 0.59 to 0.91 for different demographic groups, and 0.72 to 0.94 for the White group by Majors. For the variable "work closely related to the highest degree," the relative efficiency ranges from 0.65 to 0.93 for different demographic groups, and 0.63 to 0.81 for the White demographic group by Majors. For the selected variables, work activities in applied research, basic research, computer science, development, and professional services, the relative efficiency of the regression estimate versus the after noninterview adjusted estimate ranges from 0.81 to 1.03 for all eight demographic groups; and 0.73 to 1.08 for the White group by Majors.

VI. The Comparison of the Regression Estimate with the Current Ratio Estimate

The relative efficiencies of the regression estimate of the total to the ratio estimate for the selected variables by demographic groups are not uniformly less than one. (See Table 2) But in general, for most of the selected variables in most of the demographic groups, the regression estimate has a smaller variance than the ratio estimate. The relative efficiencies range from 0.86 to 1.03 for the selected variables in all eight demographic groups. Especially, for the survey variable " work for pay," the estimated standard error of the regression

estimate is 86 percent of the estimated standard error of the ratio estimate for the White group. On the average, the gain in efficiency of the regression estimate to the ratio estimate is about one to six percent among eight demographic groups.

For the White group by Major, the relative efficiencies range from 0.83 to 1.08. On the average, the gain in efficiency of the regression estimate to the ratio estimate is about two to six percent for the White group by Major.

VII. Summary

A regression weighting procedure is applied to the 1995 NSCG data for the selected survey variables for each demographic group and the White demographic group by Major. The 14 auxiliary control totals for each demographic group (gender, major, degree, and age group) are available for use in the regression estimate. For the White group by Major, there are 10 auxiliary control totals (gender, degree, and age group). The control totals are estimated from 1995 NSCG frames (1993 NSCG and 1993 NSRCG). We assumed that the control totals have no errors.

Assuming the 1995 NSCG frames are fixed. The variances of the regression estimate, the noninterview adjusted estimate, and the ratio adjusted estimate of the selected variables for each demographic group are estimated using the same replication methods (the successive difference replication method and the Jackknife method). The relative efficiency is used for the evaluation of the estimates. We conclude that

1. The regression estimate is better than the noninterview adjusted estimates for most of the variables considered. The average of the relative efficiencies of the regression estimate to the noninterview adjusted estimate for the selected variables range from 0.79 to 0.85 among the eight demographic groups; and from 0.80 to 0.85 for the White group by Majors.
2. The gain in efficiency of using the regression estimate over the ratio estimate ranges from 1 to 14 percent for the selected variables among the eight demographic groups; and 1 to 17 percent for the White group by Major. However, for some of the variables, there is no gain in efficiency. On average (over the selected variables), the gain is from 1 to 6 percent among the eight demographic groups; and from 2 to 6 percent for the White group by Majors.

All the empirical work of the regression estimation in this study is for the cross sectional estimation of the totals of the selected characteristics using 1995 NSCG data under the assumption that the 1995 NSCG frames are fixed, and the control totals have no errors. In practice, the 1995 NSCG frames are the 1993 NSCG and 1993 NSRCG sample respondents that have S&E degree or occupations of U.S. residents of age under 75. Further empirical research may be pursued for the longitudinal estimation of the total by using the regression weighting in multi - phase samples.

VIII. References

Brick, J.M., Broene, P., James, P., and Severynse, J. (1997), A User's Guide to WesVarPC, Version 2.1, Westat, Inc.

Cox, B.G., Cheng, J.Q. Hall, J.W., Hardy, L.P., Jang, S.D., & Riker, C.A. (1997), Sample Design For The 1995 National Survey of College Graduates, Mathematical Policy Research, Inc.

Goebel, J. J. (1976), "Application of An Iterative Regression Technique to A National Potential Cropland Survey," Proceedings of the Social Statistics Section, American Statistical Association, 350-353.

Fay, R. E. (1990), " VPLX: Variance Estimates for Complex Samples," Proceedings of the Survey Research Methods Section, American Statistical Association, 266-271.

Fay, R. E. & Train G. F. ,(1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Government Statistics Section, American Statistical Association, 154-159.

Finamore, J. M., (1997), "Revised Source and Accuracy Statement for the 1995 NSCG," Internal Census Bureau Memorandum for Chester E. Bowie from Charles H. Alexander, December 15, 1997.

Finamore, J. (1998). "Nonresponse in the 1995 National Survey of College Graduates," Demographic Statistical Methods Division, Bureau of the Census.

Fuller, W.A., Loughin, M.M.,and Baker, H. D.(1994). "Regression Weighting in the Presence of Noninterview with Application to the 1987-1988 Nationwide Food Consumption Survey," Survey Methodology, Vol. 20, No.1 ,75-85.

Huang, E.T. & Fuller, W.A. (1978), "Nonnegative Regression Estimation for Sample Survey Data," Proceedings of the Social Statistics Section, American Statistical Association 1978, 300-303.

Huang, E.T. (1978), Nonnegative Regression Estimation for Sample Survey Data. Unpublished Ph. D. thesis. Iowa State University. Ames, Iowa.

Huang, E.T. (1983), "Regression M-Weights Computer Program," Statistical Research Division, Bureau of the Census.

Moore, T. F. (1997), "Alternative Weighting for the NSCG," Internal Census Bureau Memorandum for Chester E. Bowie from Preston J. Waite, July 21, 1997.

Steinberg, J. (1992), The SUPPLY OF SCIENTISTS AND ENGINEERS IN 1989: Report of Research Concerning Additional Methods and Procedures of Weighting and Estimation for the 1989 National Survey of Natural and Social Scientists and Engineering and Advisability of Data Publication. April 6,1992. Survey Design, Inc.

Town, M. (1996), "1995 National Survey of College Graduates: The Weighting Specifications," Internal Census Bureau Memorandum for Barry G. Fink from Preston J. Waite, June 10, 1996.

Table 1 The Relative Efficiency of the Regression Estimate With the Noninterview Adjusted Estimate for the Selected Survey Variables by Demographic Group

	U.S. Born						Foreign Born		Total
	Disabled	Hispanic	White	Black	Asian	Native American	Citizen	Non Citizen	
Work for Pay	0.64	0.40	0.34	0.28	0.38	0.39	0.41	0.42	0.35
Not Looking for Work	0.59	0.91	0.79	0.87	0.87	0.83	0.85	0.88	0.80
Work Closely Related to the Highest Degree	0.93	0.75	0.72	0.65	0.79	0.77	0.77	0.74	0.73
Average of 5 Work Activities	0.94	0.93	0.94	0.91	0.95	0.92	0.94	0.93	0.94
Average of 8 Variarbles	0.86	0.84	0.82	0.79	0.85	0.82	0.84	0.84	0.82

Table 2 The Relative Efficiency of the Regression Estimate with the Ratio Estimate for the Selected Survey Variables by Demographic Group

	U.S. Born						Foreign Born		Total
	Disabled	Hispanic	Black	Asian	Native American	White	Citizen	Non Citizen	
Work for Pay	0.92	0.98	0.91	0.93	0.91	0.86	0.93	0.99	0.87
Not Looking for Work	0.90	0.91	0.92	0.90	0.89	0.86	0.91	0.98	0.87
Work Closely Related to the Highest Degree	0.99	0.99	0.97	0.99	0.95	1.00	0.98	0.97	0.98
Average of 5 Work Activities	0.98	1.00	0.99	1.00	0.95	0.98	0.99	0.99	0.99
Average of 8 Variables	0.96	0.98	0.97	0.98	0.94	0.95	0.97	0.99	0.96