NONRESPONSE ISSUES OF THE NATIONAL SURVEY OF COLLEGE GRADUATES

Antoinette Tremblay and Thomas F. Moore III
Antoinette Tremblay, U.S. Bureau of the Census, Washington, D.C. 20233

Key Words: Bias, CART, Demographic

This paper reports the general results of research undertaken by Census Bureau staff. The views are attributable to the authors and do not necessarily reflect those of the Census Bureau.

I. Synopsis

The National Survey of College Graduates (NSCG) of the National Science Foundation (NSF) collects data on U.S. scientists and engineers; it attempts to capture and measure their unusual importance to the nation's continued productivity and economic growth. The 1993 NSCG sample design reflects the efforts that were taken to reduce the effects of nonresponse. The sample came from the Sample Edited Detail File (SEDF), which contains data gathered from the 1990 Decennial Census long forms. Persons who were noninstitutionalized, U.S. residents with a Bachelor's Degree or higher, and under 76 years of age as of April 1, 1993 were eligible for the sample. The data was collected in three phases: mail, computer-assisted telephone interviewing (CATI), and personal visit.

II. Research Purpose and Scope

The purpose of this research is to assess the potential for bias, arising from the cumulative nature of nonresponse in the 1993 NSCG longitudinal survey. Across three projects, the ultimate goal is to achieve an understanding of which demographic variables, or interactions thereof, drive the phenomenon of nonresponse. By providing simple characterizations of the conditions (i.e., profiles) that determine when a sampled person is in one class (i.e., nonresponse) rather than another (i.e., response), the cost of misclassification can be reduced.

First, by comparing various 1990 census demographic variables of the 1993 NSCG nonrespondents and respondents, it can be determined if a correlation exists between certain demographic variables and nonresponse. Second, nonresponse may be correlated with frame variables or may result from survey procedures. Thus, the demographic comparison/classification analysis is repeated by reason for nonresponse. Third, a discussion of preliminary results across the data collection methods of mail, CATI, and personal visit is presented.

Each project is similar in data requirements and methodology. Final status of the NSCG and thirteen census demographic variables are obtained from the 1993 NSCG data file. In addition to considerable background and exploratory data analyses, the primary analysis is performed using the classification option of the Classification and Regression Trees (CART) statistical software of the California Statistical Software, Inc. CART is a nonparametric statistical analysis program that can automatically find hidden structures in data. Simplistically, it constructs binary decision trees from the input variables by performing various nonparametric statistical operations on a sample of the data which maximize the homogeneity of the dependent variable within each of the branches. Error improvement or reduction in error results when the use of demographic data by CART decreases the number of misclassifications from when all records are classified in the prominent category of response or nonresponse.

Since more 'traditional' analyses were desired to augment the results, conclusions, and implications obtained from the CART analysis, this research also contains some chi-square testing of the independence of various methods of classification of observed events.

III. Background Information and Data Input

Table 1 provides the final 'status codes and their descriptions for the 1993 NSCG. Of the 214,643 person records, 69.59% are defined as respondents, 21.60% as nonrespondents, and 8.81% as out-of-scope. With some additions, the variables available for classification are the same as were used in the 1993 NSCG sample selection; Table 2 shows minor regroupings of the possible values. Table 3 is a cross tabulation of these available demographic variables by the actual final status codes. Of interest are the differing percentages for PBIRTH and CTZN for emigrants, as compared to the other final status codes and other variables. Chi-square tests of the independence of the methods of classification of observed events, via contingency tables, were conducted for each of these demographic variables. With the exception of SEX, the null hypothesis that the two classifications are independent is rejected for all variables at alpha=0.10. The extremely large chi-square test statistics are caused, in part, by the large sample size, and the cells with the largest contributions to the total consistently come from the nonresponse cells.

Other background information includes response rates, where response rate=(complete+os)

divided by total. They are provided across various demographic variables in Table 4 only to depict nonresponse by different categories. Although looked at independently of the CART analyses, it is interesting to note that, overall, the 'similar' response rates across the variables will neither support nor contradict the results achieved through CART. However, the lower response rate of 65.70% for NSF GROUP=Foreign NonUS Citizen does give credibility to the analyst's choice of RACE, PBIRTH, CTZN, and the various limitation demographics as CART input variables. (NSF GROUP is never chosen by CART as a significant classification variable.) Again, this is related to Table 3's values of PBIRTH and CTZN for emigrants.

IV. Results

Results via CART and various exploratory analyses of the data are provided for each of the three nonresponse projects. Across the three projects and their various CART classification analyses, there is no demographic variable or combination of variables which have a substantial association with class membership; none are reliable predictors of response/nonresponse.

Demographic Comparison of Respondents and Nonrespondents

The out-of-scope records are omitted from this project. If no prior information is available, the lowest error rate is achieved by classifying all records as respondents. Then, a true error rate of 23.68% exists. Within rounding, the resulting error rate using CART is also 23.68%.

Knowledge of AGEGRP, CTZN, and OCCGRP provides a classification tree with an error rate lower than that of classifying all records as respondents. However, this total error rate of 23.50% is an improvement of only 0.18% (i.e., 23.68%-23.50%). CART defines a respondent as having one of three combinations of the demographic variables; AGEGRP=2,3; AGEGRP=1 CTZN=1; AGEGRP=1 CTZN=2 OCCGRP=1,2,3,4. A nonrespondent is defined by AGEGRP=1 CTZN=2 OCCGRP=5. Of the actual respondents, 98.94% are correctly classified by CART as respondents and 1.06% are incorrectly classified as nonrespondents; 4.19% of actual nonrespondents are correctly classified and 95.81% are incorrectly classified as respondents.

It is interesting to relate these results back to the independent analysis provided in Table 4. Although the response rate for the NSF GROUP involving nonUS citizenship tends to be lower than others, its value does not help considerably with the predictor via CART. Perhaps its small contribution to the reduction of the error rate may be due in part to its small part of the sample.

Chi-square tests, however, reveal that a degree of dependency may exist between response propensity and the various demographics. This methodology, although it is more familiar to most readers, does not contradict the results of CART. CART is more 'strenuous' in that it associates response/nonresponse to various demographics depending on where a majority of the records fall.

Demographic Comparison of Respondents and Nonrespondents, by Reason for Nonresponse

Looking at Table 1, reason 10c (PMR move, no forwarding) has the largest percentage of all nonresponse reasons; note that this reason is just a component of final status code 10 'Anything Else'. Refusals and persons with no Bachelor's Degree are next in priority.

Because of very few records for many of the nonresponse categories, only seven of the nonresponse reasons were used in the CART analyses: Deceased; No Bachelor's Degree; Emigrant; Refusal; PMR move, no forwarding; PMR temporarily absent; Not Located. For each of these nonresponse reasons, a first result of CART is that improvement in the total error rate is no greater than 0.76% when a three-way classification is made across respondents, the nonresponse reason, and all other nonresponse reasons. Second, when the respondents are classified against each nonresponse reason, prior demographic information is unnecessary since no classification tree is created. Regardless of the reason for nonresponse then, one can, simplistically, do no better than to designate all records as respondents. Third, the respondents were removed from the data set and each of the seven nonresponse reasons were classified against all others. Prior demographic information is again unnecessary. Lastly, a classification across all nonresponse reasons was done. The operational implications are vague, if any, but the result is of interest. Since the nonresponse reason of "PMR move, no forwarding" is the largest, CART analysis indicates that all nonrespondents should be classified as having this reason, with a 7.80% improvement in total error rate and with all thirteen demographic variables entering into the classification.

Differences in Results Across Data Collection Modes

It was hoped that results from this project would help answer some of the following questions: Is it worth doing only one or two interview modes for various subsets of the sample? For example, perhaps mail could be eliminated for some 'profiles' that exhibit strong dependence with nonresponse, and the data could then be collected initially via CATI or personal visit.

It is possible to compare respondents and nonrespondents across the three data collection modes of mail, CATI, and personal visit since the NSCG data

set provides the 'intermediate' status codes for each record, after each data collection effort. (Hereafter, these intermediate status codes are designated outcome codes). Provided as background, the percent of NSCG person records undergoing each type (and combination) of collection mode is given in Table 5.

CART analyses of respondents vs. nonrespondents are then conducted on five data sets which differ in terms of which outcome code is used as the final status code. As in the first project, the out-of-scope records are omitted here. It is assumed that the progression of data collection is mail, CATI and personal visit. The first data set consists of just those records which have a mail outcome code; the second data set consists of just those records which have a CATI outcome code; likewise, the third consists of those records having an outcome code resulting from a personal visit. Addressing just CATI, there are 135,097 records which never underwent this data collection method; therefore, the fourth data set substitutes the mail outcome code, if present, for these records. Addressing just personal visits, the fifth data set first substitutes any present CATI outcome code for records with missing personal visit outcome code, and then substitutes the mail outcomes code for those records which did not undergo either CATI or personal visit. Table 6 reveals that results for this project are more noticeable than the other two projects. When looking at mail vs. CATI vs. personal visit, mail is the only mode which reveals any possibility for error improvement when information on various demographics is available. It exhibits an error improvement of 6.74%; however, the profiles of these CART respondents and nonrespondents, it is felt, are too cumbersome to incorporate operationally. For the other CART analysis which exhibits a non-zero error improvement, all nonrespondents should be classified as nonrespondents if no prior information is available; however, prior information does decrease the error rate to 2.02%. For the remaining three CART analyses, the best that one can do is to classify all nonresponse records to whichever outcome (response or nonresponse) occurs more frequently.

A complementary and vital issue in this discussion is the tie-in of cost per completed interview across the three data collection modes. Suppose the cost and response rates were known each for mail and CATI; then one could perform some desirable future tradeoffs. These issues should definitely should be researched further since the implications may prove to be rewarding.

V. Final Remarks

An extensive amount of background, exploratory data analysis, chi-square testing, and CART classification analysis was performed in this research. Simplistically, the goals were to provide an interpretable picture of a structure for the 1993 NSCG data and to determine if any of thirteen 1990 Decennial Census demographic variables could reliably distinguish the survey's respondents from their nonrespondents. The results could not provide consistent characterizations of the conditions that determine when a sample person is a respondent rather than a nonrespondent. In terms of reasons for nonresponse, the implications of the large majority of records having the nonresponse reason of "PMR move, no forwarding" should be explored for the goal of nonresponse reduction. Also, the group that stands out with the lowest response rate is the Foreign born, nonUS citizens. In terms of modes of data collection, mail is the only mode which reveals any possibility for error improvement (6.74%), but the profiles are far too complicated. The conclusions and implications of the results presented need to be considered more thoroughly.

Further research in this arena is recommended. Perhaps complementary to current research by Groves and Couper in "Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys," topics could include a CART classification analysis that attempts to distinguish refusals from noncontacts. Also, a considerable amount of effort into obtaining and utilizing cost data across the data collection modes is recommended.

VI. Supporting Materials

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification And Regression Trees*, Belmont, California: Wadsworth International Group.

California Statistical Software, Inc. (1985), "An Introduction to CART Methodology."

California Statistical Software, Inc. (1993), "Using the CART Programs, Version 1.3."

Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: John Wiley & Sons.

Groves, R. M. and Couper, M. P., "Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys."

U.S. Bureau of the Census (1993), "Sample Selection Specifications for the 1993 National Survey of College Graduates (NSCG) - Revised," memorandum from Preston Jay Waite to Thomas C. Walsh, June 15, 1993.

Table 1. NSCG Final Status Codes

Final Status Code		Description	Total	%
Response	1	Complete	149377	69.59
Out-of-Scope[1]	2	Age Over 75	211	0.10
	3	Deceased	2407	1.12
	4	No Bachelor's Degree	14232	6.63
	5	Emigrant	1904	0.89
	6	Institutionalized	159	0.07
Nonresponse	7	Ill	1833	0.85
	8	Refusal	15082	7.03
	9	Incomplete	625	0.29
	10	Anything Else	28813	13.43
	10a	PMR with correction (move)	298	0.14
	10b	PMR Jeffersonville correction	4	0.00
	10c	PMR move, no forwarding	19460	9.07
	10d	PMR forwarding expired	14	0.00
	10e	PMR temporarily absent,...	2090	0.97
	10f	Not located	5300	2.47
	10g	Wrong person	1333	0.62
	10h	Foreign address, APO	24	0.01
	10i	Not received	290	0.14
Total			214643	100.00

Table 2. Demographic Variables Available for Classification

Demographic Characteristic (variable name)	Values
Age Group (AGEGRP)	1=[16,29]; 2=[30,59]; 3=60+
Sex (SEX)	1=Male; 2=Female
Race (RACE)	1=White
	2=Black
	3=Native American
	4=Asian/Pacific Islander
	5=Other
Spanish/Hispanic Origin (ORIGIN)	1=No; 2=Yes
Place of Birth (PBIRTH)	1=US or Outlying Area; 2=Other
Citizenship (CTZN)	1=Yes; 2=No
Highest Education Degree (EDUC)	1=Bachelor's or Professional
	2=Master's
	3=Doctorate
Occupation Group (OCCGRP)	1=Physics/Life/Biology Scientists
	2=Math/Computer Scientists
	3=Social Scientists
	4=Engineers, Architects, Surveyors
	5=Other
Mobility Limitation Status (MOLIMT)	1=Yes; 2=No
Personal Care Limitation (PCLIMT)	1=Yes;2=No
Work Limitation Status (WRKLIMT)	1=Yes; 2=No
Work Prevention Status (WRKPVT)	1=Yes; 2=No
Metropolitan Statistical Area (MSA)	1=Yes; 2=No; 9=Missing

[1] Codes 2, 3, and 4 are permanently out-of-scope; codes 5 and 6 are temporarily out-of-scope.

Table 3. Classification Variables by Final Status Code (%)

Demographic Variable	Final Status Code										Total
	1	2	3	4	5	6	7	8	9	10	
AGEGRP=1	18	10	5	26	32	16	19	15	17	39	21
AGEGRP=2	72	27	50	64	65	50	65	76	64	57	69
AGEGRP=3	10	62	45	11	4	33	16	10	19	4	10
SEX=1	59	51	73	53	64	69	66	63	53	56	59
SEX=2	41	49	27	47	36	31	34	37	47	44	41
RACE=1	79	80	83	68	56	75	71	79	68	64	76
RACE=2	9	14	10	15	4	16	12	10	11	17	10
RACE=3	1	<1	1	1	<1	3	1	1	1	1	1
RACE=4	10	5	5	11	37	2	14	9	18	14	11
RACE=5	2	1	1	4	4	4	2	1	2	4	2
ORIGIN=1	94	94	95	87	87	90	93	95	89	88	93
ORIGIN=2	6	6	5	13	13	10	7	5	11	12	7
PBIRTH=1	83	80	87	73	27	89	70	81	65	70	80
PBIRTH=2	17	20	13	27	73	11	30	19	35	30	20
CTZN=1	93	91	96	86	42	94	87	93	82	81	91
CTZN=2	7	9	4	14	58	6	13	7	18	19	9
EDUC=1	69	73	71	85	61	73	73	74	78	75	71
EDUC=2	26	22	24	13	28	21	22	22	17	20	24
EDUC=3	5	4	5	2	11	6	5	4	4	4	5
OCCGRP=1	4	2	3	2	4	0	3	3	2	3	3
OCCGRP=2	5	1	3	2	4	3	4	5	3	4	5
OCCGRP=3	3	2	2	1	3	1	3	3	1	3	3
OCCGRP=4	13	4	10	7	10	8	9	12	8	9	12
OCCGRP=5	76	91	83	88	79	89	82	78	86	81	78
MOLIMT=1	1	9	17	3	<1	30	6	2	2	1	2
MOLIMT=2	99	91	83	97	100	70	94	98	98	99	99
PCLIMT=1	2	10	12	5	2	24	6	3	4	3	3
PCLIMT=2	98	90	88	95	98	76	94	97	96	97	97
WRKLIMT=1	6	28	41	10	2	50	14	7	10	6	7
WRKLIMT=2	94	72	59	90	98	50	86	93	90	94	93
WRKPVT=1	2	22	24	4	1	39	8	2	5	2	2
WRKPVT=2	98	78	76	97	99	61	92	98	95	98	98
MSA=1	11	14	13	11	8	11	9	8	10	8	10
MSA=2	89	85	87	88	92	88	91	91	89	92	90
MSA=9	1	1	1	1	<1	1	1	1	<1	1	1
Total	70	<1	1	7	1	<1	1	7	<1	13	

Table 4. Response Rates

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
AGEGRP=1	26918	3790	628	13908	45244	69.26
AGEGRP=2	107056	10326	1308	29442	148132	80.12
AGEGRP=3	15403	2734	127	3003	21267	85.88
SEX=1	87922	9373	1336	27258	125889	78.35
SEX=2	61455	7477	727	19095	88754	78.49
OCCGRP=1	5339	295	81	1451	7166	79.75
OCCGRP=2	7346	415	73	2019	9853	79.51
OCCGRP=3	4216	249	61	1366	5892	76.82
OCCGRP=4	18961	1206	198	4479	24844	81.97
OCCGRP=5	113515	14685	1650	37038	166888	77.81
EDUC=1	102604	13982	1268	34689	152543	77.26
EDUC=2	38673	2404	572	9644	51293	81.20
EDUC=3	8100	464	223	2020	10807	81.31
NSF GROUP:						
Disabled	11072	2636	119	3613	17440	79.28
Hispanic	5914	959	64	2332	9269	74.84
White/Other	90751	6784	403	21588	119526	81.94
Black	10877	1756	42	5183	17858	70.98
Asian/Pacific I	3511	307	31	883	4732	81.34
Natv American	1166	191	5	478	1840	74.02
Frg US Ctz	16051	2148	285	5374	23858	77.48
Frg NonUS Ctz	10035	2069	1114	6902	20120	65.70
Total	149377	16850	2063	46353	214643	78.40

Table 5. Distribution by Data Collection Mode

Data Collection Mode	Records	Percentage
Mail only	117531	54.76
CATI only	3709	1.73
Personal Visit only	855	0.40
Mail and CATI only	39374	18.34
Mail and Personal Visit only	16711	7.78
CATI and Personal Visit only	1686	0.79
Mail and CATI and Personal Visit	34777	16.20
Total	214643	100.00

Table 6. Results of CART Classifications, by Mode of Data Collection (%)

Data Set	No Prior Info Error Rate	CART Error Rate	Error Improvement
Mail	44.15	37.41	6.74
CATI	30.57	30.57	0
Personal Visit	36.03	36.03	0
CATI, w/Mail replacements	33.58	31.56	2.02
Personal Visit, w/CATI & Mail repl.	24.15	24.15	0