**16. Weighting the Personnel Survey: One Solution**

The project assigned in Chapter 12 was to compute a set of weights for a survey of members of the military reserves. A stratified simple random sample of personnel was selected and queried about satisfaction with their jobs. The project provides an opportunity to put into practice the techniques covered in Chapters 13-15. Completing the project requires calculation of base weights, an adjustment to account for cases whose eligibility status is unknown, an adjustment for nonresponse, and calibration to some finite population totals. There are several practical problems to be solved, including selecting a particular method of nonresponse adjustment, deciding how to use the population counts that are available, and determining how to handle missing values in both the sample cases and the population counts.

Although this chapter is not written as a formal report to be delivered to a client as was requested in the Chapter 12 assignment, we want to again emphasize the importance of good documentation. Clear documentation of all weighting steps is critical for several reasons. It may be necessary to repeat some or all steps at a later time. For example, errors may be discovered in some details of calculations, or problems may be found in one of the input data sets. If a survey will be repeated at a later date, a well written weighting report can guide the work in the next survey. Very detailed specification memos, like the ones described in Chapter 18, will remove any doubt about what should be done and can lead to reduced costs if the survey is repeated at a later date.

The R code for the solution to this project is in the files,

16.1 Solution bwt-unknown adj.R

16.2 Solution NR adj.R

16.3 Solution calibration adj.R

16.4 Example tabulations.R

all of which are on the book’s website.

**16.1 The Data Files**

Two data files were provided for the project. One file (SOFR.sas7bdat/SOFR.xpt) contained records for all 71,701 sample members who were initially selected. The file includes the 19 variables shown in Table 16.1. The fields include identification number, final respondent status code, stratum identifier, stratum sample count and population count, frame variables (gender, pay grade, race, etc.), and respondents’ answers to key questions. The variable RESPSTAT for final respondent status code has information about the eligibility and the response status for each sample member.

The fields NSAMP and NSTRAT contain the number of cases in the sample and in the frame for the stratum to which a person belongs. The values are the same for all records for persons in a given stratum. Based on inspecting the file of sample persons, there were 404 strata, defined by combinations of branch of the service, race-ethnicity, gender, and paygrade.

**Table 16.1.** Contents of data file SOFR.sas7bdat

|  |  |  |
| --- | --- | --- |
| **#** | **Variable** | **Label** |
| 1 | REC\_ID | Unique Record Identification Number |
| 2 | RESPSTAT | Final Respondent Status Code |
| 3 | SRMARST | What is your marital status? |
| 4 | RA006A | Taking all things into consideration, how satisfied are you, in general, with each of the following aspects of being in the National Guard/Reserve? Your total compensation (i.e., base pay, allowances, and bonuses) |
| 5 | RA006B | Taking all things into consideration, how satisfied are you, in general, with each of the following aspects of being in the National Guard/Reserve? The type of work you do in your military job |
| 6 | RA008 | Suppose that you have to decide whether to continue to participate in the National Guard/Reserve. Assuming you could stay, how likely is it that you would choose to do so? |
| 7 | RA115 | Overall, how well prepared are you to perform your wartime job? |
| 8 | RA118 | Overall, how would you rate the current level of stress in your personal life? |
| 9 | SRED | What is the highest degree or level of school that you have completed? Mark the one answer that describes the highest grade or degree that you have completed. |
| 10 | RA112RA | Average: In past 12 months, how many days did you spend in a compensated Reserve/Guard status? |
| 11 | XSRRCR | **Recoded: Imputed Service** |
| 12 | XACT2R | **Recoded: Imputed Activated 30 Days - 3 Level** In the last 24 months were you ever activated longer than 30 consecutive days? |
| 13 | XRETH4R | **Recoded: Imputed Race/Ethnicity - 2 Level** |
| 14 | XSEXR | **Recoded: Imputed Gender** |
| 15 | XCPAY1R | **Recoded: Imputed Paygrade Group 1** |
| 16 | NSAMP | Stratum Sample Count |
| 17 | NSTRAT | Stratum Population Count |
| 18 | V\_STRAT | Variance estimation stratum |
| 19 | STRATUM | Design stratum |

As shown in Table 16.2, the other data file (RCCPDS57.sas7bdat/RCCPDS57.xpt) has population counts for seven frame variables (branch of the service, gender, pay grade, race-ethnicity, education, marital status, and whether a person had been called to active service more than 30 consecutive days in the last 24 months.). These frame variables have different names than in the sample data file, but the alternative names are indicated in the labels. Population counts are provided in variable COUNT.

**Table 16.2.** Contents of data file RCCPDS57.sas7bdat

|  |  |  |
| --- | --- | --- |
| **#** | **Variable** | **Label** |
| 1 | SERVICE | (XSRRCR) Branch of Military Service |
| 2 | GENDER | (XSEXR) Gender |
| 3 | PG\_GROUP | (XCPAY1R) Pay grade group |
| 4 | RACETH | (XRETH4R) Race/Ethnicity |
| 5 | EDUCCAT | (SRED) Highest degree/level of school completed |
| 6 | MARIT | (SRMARST) Current marital status |
| 7 | ACTIVATD | (XACT2R) Activated more than 30 consecutive days or less in last 24 months |
| 8 | COUNT | Person count |

**16.2 Base Weights**

Base weights can be computed as soon as the sample is selected. We do not need to know the dispositions of any of the sample cases because the base weights in this survey depend only on the frame counts and the sample sizes in each of the design strata. Since a stratified simple random sample was selected, the selection probability of each person *i* in stratum *h* was  where

 = number of persons sampled from stratum *h*;

 = number of persons on the frame in stratum *h*.

The base weight for person *hi* is the inverse of the selection probability: . This is computed as NSTRAT/NSAMP. The sum of the base weights is 870,373, which is exactly equal to the count of the persons on the frame since the sample is *stsrs*.

**16.3 Disposition Codes and Mapping into Weighting Categories**

Table 16.3 gives counts of persons by the disposition codes in the RESPSTAT field. These codes are specific to the survey of Reserve personnel, as is apparent from some of the categories. For example, code 22 (No return—separated/retired) would probably not be used in surveys of most other populations. Because there was a time lapse between the time the sample was selected and the data were collected, the status of some persons changed. This is the reason for having codes for retirees, deceased, incarcerated, etc. Addresses for some personnel are out-of-date leading to the inability of the postal service to deliver the survey (code 27). To compute weights, the disposition codes need to be mapped into the groups:

|  |  |
| --- | --- |
| ER | Eligible respondents |
| ENR | Eligible nonrespondents |
| IN | Known ineligibles |
| UNK | Unknown eligibility |

To compute the various AAPOR response rates described in Chapter 18, the disposition codes are mapped to a slightly different set of categories:

|  |  |
| --- | --- |
| I | Complete interview |
| P | Partially complete interview |
| R | Refusal/Break-off |
| NE | Not eligible |
| U | Unknown eligibility |
| O | Other eligible noninterview |

The mappings we used for both the weighting and AAPOR categories are shown in Table 16.4. A number of decisions have to be made about how to map the dispositions. Some choices are obvious, like mapping code 1 (Questionnaire Returned – Completed) to ER and I. Others are less so, like codes 5 (Questionnaire Returned – Blank), 25 (No Return – Other), 27 Postal non-delivery), and 29 (Not Locatable). Unless more is known about such cases, a conservative approach would be to consider the eligibility of these persons as unknown, which we did in Table 16.4. Since there is disposition code 26 (No Return — eligible based on administrative records), it is apparent that efforts were made to match the sample file against personnel records. Consequently, the alternative argument could be made that persons in codes 5, 25, 27, and 29 are ineligible. Clearly, there is some subjectivity in the mapping.

**Table 16.3.** Counts for each final respondent status code

|  |  |
| --- | --- |
| Respondent status (as stored in RESPSTAT variable) | Count |
| 1 = Questionnaire Returned – Completed | 25,539 |
| 2 = Questionnaire Returned – (Sufficient) Partial Complete | 20 |
| 3 = Questionnaire Returned – (Insufficient) Partial Complete | 524 |
| 4 = Questionnaire Returned – Ineligible | 503 |
| 5 = Questionnaire Returned – Blank | 97 |
| 18 = No Return – Deceased | 9 |
| 19 = No Return – Incarcerated | 2 |
| 22 = No Return – Separated/Retired | 35 |
| 23 = No Return – Active Refusal | 193 |
| 25 = No Return – Other | 8 |
| 26 = No Return — eligible based on administrative records | 39,872 |
| 27 = Postal Non-delivery | 1,339 |
| 29 = Not Locatable | 6 |
| 35 = Ineligible – No Questionnaire Sent | 3,554 |
| Total | 71,701 |

**Table 16.4**. Mapping of disposition codes into collapsed respondent statuses

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Response status (as stored in RESPSTAT)** | **Weighting code** | **Weighting category** | **AAPOR code** | **AAPOR description** |
| 1 = Questionnaire Returned – Completed | ER | Eligible respondent | I | Complete interview |
| 2 = Questionnaire Returned – (Sufficient) Partial Complete | ER | Eligible respondent | P | Partially complete interview |
| 3 = Questionnaire Returned – (Insufficient) Partial Complete | ENR | Eligible nonrespondent | R | Refusal/Break-off |
| 4 = Questionnaire Returned – Ineligible | IN | Ineligible | NE | Not eligible |
| 5 = Questionnaire Returned – Blank | UNK | Unknown eligibility | U | Unknown eligibility |
| 18 = No Return – Deceased | IN | Ineligible | NE | Not eligible |
| 19 = No Return – Incarcerated | IN | Ineligible | NE | Not eligible |
| 22 = No Return – Separated/Retired | IN | Ineligible | NE | Not eligible |
| 23 = No Return – Active Refusal | ENR | Eligible nonrespondent | R | Refusal/Break-off |
| 25 = No Return – Other | UNK | Unknown eligibility | U | Unknown eligibility |
| 26 = No Return — eligible based on administrative records | ENR | Eligible nonrespondent | O | Other eligible noninterview |
| 27 = Postal Non-delivery | UNK | Unknown eligibility | U | Unknown eligibility |
| 29 = Not Locatable | UNK | Unknown eligibility | U | Unknown eligibility |
| 35 = Ineligible – No Questionnaire Sent | IN | Ineligible | NE | Not eligible |

Table 16.5 shows the counts of cases in the weighting and AAPOR categories. Judging from the counts, unknown eligibility is a minor problem. On the other hand, the response rate is well under 50%. Thus, concentrating efforts on the nonresponse adjustment is prudent in this sample.

**Table 16.5**. Counts for each weighting and AAPOR category

|  |  |  |  |
| --- | --- | --- | --- |
| Disposition | Indicator | Count | Percent |
| Weighting category (disposition codes) | |  |  |
| Eligible respondent (1,2) | ER | 25,559 | 35.6 |
| Eligible nonrespondent (3,23,26) | NR | 40,589 | 56.6 |
| Known ineligible (4,18,19,22,35) | IN | 4,103 | 5.7 |
| Unknown eligibility (5,25,27,29) | UNK | 1,450 | 2.0 |
| Total |  | 71,701 | 100.0 |
|  |  |  |  |
| AAPOR category (disposition codes) |  |  |  |
| Complete (1) | I | 25,539 | 35.6 |
| Partial (2) | P | 20 | 0.03 |
| Refusal/Break-off (3,23) | R | 717 | 1.0 |
| Other eligible noninterview (26) | O | 39,872 | 55.6 |
| Not eligible (4, 18, 19, 22, 35) | NE | 4,103 | 5.7 |
| Unknown eligibility (5, 25, 27, 29) | U | 1,450 | 2.0 |
| Total |  | 71,701 | 100.0 |

Chapter 18 reviewed various outcome rates that may be computed in a survey. As illustrations, we compute *RR*1and *RR*4 which are defined as:

 and

 where



is the proportion of unknowns that are allocated to being eligible. In this sample, *e* = 0.941, *RR*1= 37.78%, and *RR*4= 37.83%. Since the number of unknowns is a small part of the full sample, the values of these two response rates are virtually the same.

**16.4 Adjustment for Unknown Eligibility**

Using the base weights, we can estimate the numbers of persons on the frame that are in the weighting categories, ER, ENR, IN, and UNK:

|  |  |  |  |
| --- | --- | --- | --- |
| Weighting category |  | Estimated count | Percentage |
| Eligible respondent | ER | 320,677 | 36.8% |
| Eligible nonrespondent | NR | 474,675 | 54.5% |
| Known ineligible | IN | 55,770 | 6.4% |
| Unknown eligibility | UNK | 19,251 | 2.2% |
| Total |  | 870,373 | 100.0% |

The estimated population counts are distributed in about the same way as the unweighted counts in Table 16.5. Since only 2.2% of the frame is estimated to be unknowns, we will make one overall adjustment, which, using the notation from section 14.2, is equal to

.

The adjustment is made in the file 16.1 Solution bwt-unknown adj.R.

**16.5 Variables Available for Nonresponse Adjustment**

There are four variables that have non-missing data for both the sample respondents and nonrespondents: branch of the service, race-ethnicity, gender, and paygrade. These are the same variables that were used in defining design strata. The other personal characteristics—education, marital status, and whether a person spent more than 30 consecutive days on active duty in the last 24 months—are each missing for almost all nonrespondents. Table 16.6a shows sample counts of responding and nonresponding persons for each of the variables that we can use for nonresponse adjustment; Table 16.6b shows similar counts for the other three demographic variables that are available mainly for respondents. The two tables also show population counts from the RCCPDS57.XPT file.

The file from which population counts were made had some missing data for each variable, other than pay group. For example, branch of the service in Table 16.6a was missing for 291 persons; race-ethnicity was missing for 602 persons. Later, in section 16.\_, when we calibrate to the population counts, imputations will have to be made for those missing values.

**Table 16.6a**. Sample counts of respondents and nonrespondents and population counts for the four variables with no missing data for sample persons.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| (Code value) Variable | Nonrespondent | | Respondent | |  | Population controls (before imputation) | |
|  | *n* | % | *n* | % | Total | *N* | % |
| **Service** |  |  |  |  |  |  |  |
| (1) Army National Guard | 10,060 | 65.0% | 5,424 | 35.0% | 15,484 | 322,053 | 40.2% |
| (2) Army Reserve | 8,398 | 61.9% | 5,179 | 38.1% | 13,577 | 190,235 | 23.7% |
| (3) Naval Reserve | 4,686 | 56.4% | 3,617 | 43.6% | 8,303 | 77,022 | 9.6% |
| (4) Marine Corps Reserve | 7,869 | 70.6% | 3,283 | 29.4% | 11,152 | 36,094 | 4.5% |
| (5) Air National Guard | 4,855 | 53.6% | 4,207 | 46.4% | 9,062 | 105,092 | 13.1% |
| (6) Air Force Reserve | 4,721 | 55.1% | 3,849 | 44.9% | 8,570 | 71,022 | 8.9% |
| Missing | — | — | — | — | — | 291 | 0.04% |
| **Race-ethnicity** |  |  |  |  |  |  |  |
| (1) NonHispanic White | 20,625 | 55.1% | 16,833 | 44.9% | 37,458 | 540,473 | 67.4% |
| (2) Total Minority | 19,964 | 69.6% | 8,726 | 30.4% | 28,690 | 260,734 | 32.5% |
| Missing | — | — | — | — | — | 602 | 0.1% |
| **Gender** |  |  |  |  |  |  |  |
| (1) Male | 34,100 | 61.9% | 21,007 | 38.1% | 55,107 | 663,122 | 82.7% |
| (2) Female | 6,489 | 58.8% | 4,552 | 41.2% | 11,041 | 138,574 | 17.3% |
| Missing | — | — | — | — | — | 113 | 0.01% |
| **Pay Group** |  |  |  |  |  |  |  |
| (1) E1 - E3 | 7,026 | 82.5% | 1,494 | 17.5% | 8,520 | 112,244 | 14.0% |
| (2) E4 | 12,936 | 75.8% | 4,125 | 24.2% | 17,061 | 198,048 | 24.7% |
| (3) E5 - E6 | 10,146 | 64.2% | 5,653 | 35.8% | 15,799 | 265,388 | 33.1% |
| (4) E7 - E9 | 2,810 | 47.1% | 3,162 | 52.9% | 5,972 | 110,397 | 13.8% |
| (5) W1 - W5 | 987 | 42.1% | 1,356 | 57.9% | 2,343 | 10,948 | 1.4% |
| (6) O1 - O3 | 3,185 | 45.7% | 3,783 | 54.3% | 6,968 | 41,176 | 5.1% |
| (7) O4 - O6 | 3,499 | 36.9% | 5,986 | 63.1% | 9,485 | 63,608 | 7.9% |
| Missing | — | — | — | — | — | — | — |
|  |  |  |  |  |  |  |  |
| Grand totals | 40,589 | 61.4% | 25,559 | 38.6% | 66,148 | 801,809 | 100.0% |

**Table 16.6b**. Sample counts of respondents and nonrespondents and population counts for education level, marital status, and activation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| (Code value) Variable | Nonrespondent | | Respondent | |  | Population controls (before imputation) | |
|  | *n* | % | *n* | % | Total | *N* | % |
| **Education** |  |  |  |  |  |  |  |
| (1) 12 years or less of school (no diploma) | 0 | 0.0% | 146 | 100.0% | 146 | 10,819 | 1.3% |
| (2) High school graduate--high school diploma or equivalent (e.g., GED) | 0 | 0.0% | 2,059 | 100.0% | 2,059 | 116,933 | 14.6% |
| (3) Some college credit, but less than 1 year | 1 | 0.0% | 2,465 | 100.0% | 2,466 | 113,512 | 14.2% |
| (4) 1 or more years of college, no degree | 2 | 0.0% | 4,967 | 100.0% | 4,969 | 223,581 | 27.9% |
| (5) Associate's degree | 1 | 0.0% | 2,399 | 100.0% | 2,400 | 96,073 | 12.0% |
| (6) Bachelor's degree | 8 | 0.1% | 7,750 | 99.9% | 7,758 | 147,450 | 18.4% |
| (7) Master's, doctoral or professional school degree | 1 | 0.0% | 4,912 | 100.0% | 4,913 | 66,614 | 8.3% |
| Missing | 40,576 | 97.9% | 861 | 2.1% | 41,437 | 26,827 | 3.3% |
| **Marital status** |  |  |  |  |  |  |  |
| (1) Married | 233 | 1.4% | 16,934 | 98.6% | 17,167 | 455,603 | 56.8% |
| (2) Separated | 6 | 1.5% | 397 | 98.5% | 403 | 11,748 | 1.5% |
| (3) Divorced | 40 | 1.6% | 2,538 | 98.4% | 2,578 | 75,025 | 9.4% |
| (4) Widowed | 0 | 0.0% | 75 | 100.0% | 75 | 3,324 | 0.4% |
| (5) Never Married | 157 | 2.7% | 5,577 | 97.3% | 5,734 | 254,468 | 31.7% |
| Missing | 40153 | 99.9% | 38 | 0.1% | 40,191 | 1,641 | 0.2% |
| **Activated more than 30 days** |  |  |  |  |  |  |  |
| (1) Activated <= 30 Days | 7 | 1.1% | 611 | 98.9% | 618 | 37,171 | 4.6% |
| (2) Activated > 30 Days | 148 | 1.1% | 12,912 | 98.9% | 13,060 | 250,808 | 31.3% |
| (3) Not activated | 24 | 0.2% | 11,814 | 99.8% | 11,838 | 508,083 | 63.4% |
| Missing | 40,410 | 99.5% | 222 | 0.5% | 40,632 | 5,747 | 0.7% |
|  |  |  |  |  |  |  |  |
| Grand totals | 40,589 | 61.4% | 25,559 | 38.6% | 66,148 | 801,809 | 100.0% |

**16.6 Nonresponse Adjustments**

Two options for nonresponse adjustment that we covered in Chapter 13 are to use estimated response propensities and cells formed with a regression tree. Both alternatives are examined in this section.

**Propensity models**

First, we will examine the option of creating classes based on estimated response probabilities or propensities. A model with main effects and all two-way interactions, using the four available variables was fitted without using survey weights. The R code is shown below and is in the file 16.2 Solution NR adj.R. The variable resp is 1 for respondents and 0 for nonrespondents.

glm.logit2 <- glm(resp ~ as.factor(xsrrcr)\*as.factor(xreth4r)

+ as.factor(xsrrcr)\*as.factor(xsexr)

+ as.factor(xsrrcr)\*as.factor(xcpay1r)

+ as.factor(xreth4r)\*as.factor(xsexr)

+ as.factor(xreth4r)\*as.factor(xcpay1r)

+ as.factor(xsexr)\*as.factor(xcpay1r),

family=binomial(link = "logit"),

data = sofr.d1.elig)

anova(glm.logit2, test="Chisq")

The data set sofr.d1.elig is a subset of sofr.sas7bdat that contains only the 66,148 eligible respondents and nonrespondents. The output from the anova command is shown below.

Df Deviance Resid. Df Resid. Dev P(>|Chi|)

as.factor(xsrrcr) 5 951.3 66142 87304 < 2.2e-16 \*\*\*

as.factor(xreth4r) 1 1376.9 66141 85927 < 2.2e-16 \*\*\*

as.factor(xsexr) 1 13.2 66140 85914 0.0002764 \*\*\*

as.factor(xcpay1r) 6 6081.2 66134 79833 < 2.2e-16 \*\*\*

as.factor(xsrrcr):as.factor(xreth4r) 5 71.0 66129 79762 6.379e-14 \*\*\*

as.factor(xsrrcr):as.factor(xsexr) 5 15.9 66124 79746 0.0070584 \*\*

as.factor(xsrrcr):as.factor(xcpay1r) 30 209.9 66094 79536 < 2.2e-16 \*\*\*

as.factor(xreth4r):as.factor(xsexr) 1 31.2 66093 79505 2.291e-08 \*\*\*

as.factor(xreth4r):as.factor(xcpay1r) 6 7.2 66087 79498 0.3004589

as.factor(xsexr):as.factor(xcpay1r) 6 36.8 66081 79461 1.937e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

All main effects and interactions are highly significant except for the xreth4r\*xcpay1r interaction. In a survey-weighted regression the same factors and interactions were significant. With such a large sample size, we could probably find some significant three-way interactions also. But, for this project we will not attempt to extend the model above.

Given predicted response probabilities from this model, we can create classes based on their quantiles. Table 16.7 shows the ranges of propensities and counts of persons in each class when 5 and 10 classes are created. Notice that the counts of persons in each class are not equal. Since the model uses only factors as predictors, there are many ties among the estimated propensities, leading to uneven divisions among the classes. Using 10 classes does seem to distinguish better among different rates than does 5 classes. The estimated rates within classes in the last five columns of Table 16.7 are fairly similar regardless of the method of calculation.

**Table 16.7** Ranges of estimated response propensities for 5 and 10 classes along with five estimates of response propensity within each class.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Range of estimated propensities | Number of sample persons | Unweighted mean | Weighted mean | Unweighted response rate | Weighted response rate | Median response propensity |
| 5 classes |  |  |  |  |  |  |
| [0.098,0.213] | 13,568 | 0.160 | 0.172 | 0.160 | 0.201 | 0.166 |
| (0.213,0.300] | 13,539 | 0.262 | 0.251 | 0.260 | 0.267 | 0.264 |
| (0.300,0.446] | 13,077 | 0.375 | 0.387 | 0.380 | 0.416 | 0.375 |
| (0.446,0.569] | 13,213 | 0.520 | 0.509 | 0.521 | 0.517 | 0.520 |
| (0.569,0.735] | 12,751 | 0.631 | 0.635 | 0.629 | 0.651 | 0.628 |
|  |  |  |  |  |  |  |
| 10 classes |  |  |  |  |  |  |
| [0.098,0.166] | 8,218 | 0.140 | 0.147 | 0.141 | 0.177 | 0.137 |
| (0.166,0.213] | 5,350 | 0.191 | 0.198 | 0.190 | 0.226 | 0.190 |
| (0.213,0.264] | 7,600 | 0.244 | 0.242 | 0.236 | 0.256 | 0.238 |
| (0.264,0.300] | 5,939 | 0.285 | 0.284 | 0.290 | 0.313 | 0.281 |
| (0.300,0.360] | 6,053 | 0.336 | 0.335 | 0.346 | 0.350 | 0.326 |
| (0.360,0.446] | 7,024 | 0.409 | 0.403 | 0.409 | 0.437 | 0.398 |
| (0.446,0.519] | 6,182 | 0.481 | 0.472 | 0.478 | 0.494 | 0.484 |
| (0.519,0.569] | 7,031 | 0.555 | 0.551 | 0.558 | 0.542 | 0.555 |
| (0.569,0.627] | 6,136 | 0.595 | 0.599 | 0.588 | 0.611 | 0.591 |
| (0.627,0.735] | 6,615 | 0.665 | 0.667 | 0.667 | 0.687 | 0.649 |
|  |  |  |  |  |  |  |
| Total | 66,148 |  |  |  |  |  |



**Figure 16.1** Boxplots of estimated response propensities grouped into 5 and 10 classes. Propensity model estimated based on the four variables available for respondents and nonrespondents using a model with main effects and all two-way interactions. A dot marks the average propensity in each class.

Figure 16.1 shows boxplots of the estimated propensities from the model for the 5 and 10 class breakdowns. The ranges are fairly wide in each of the 5 classes, but noticeably less within each of the 10 classes. As a further diagnostic, we can check whether balance was achieved for the covariates in the 10-class breakdown. The following R code creates an indicator for whether a person is in the Army National Guard and checks balance.

v1 <- rep(0,nrow(sofr.d1.elig))

v1 <- sofr.d1.elig$xsrrcr == 1 # Army National Guard

chk <- glm(v1 ~ as.factor(p.class.10) + as.factor(resp) + as.factor(p.class.10)\*as.factor(resp),

family=binomial(link = "logit"),

data = sofr.d1.elig)

anova(chk, test="Chisq")

The output of the anova statement is

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Df | Deviance | Resid. Df | Resid. Dev | P(>|Chi|) |  |
| as.factor(p.class.10) | 9 | 12764.8 | 66138 | 59225 | < 2.2e-16 | \*\*\* |
| as.factor(resp) | 1 | 8.9 | 66137 | 59217 | 0.00288 | \*\* |
| as.factor(p.class.10):  as.factor(resp) | 9 | 24.5 | 66128 | 59192 | 0.00364 | \*\* |

Similar checks (not shown here) reveal that the interaction term is significant when predicting whether a person is in the Army Reserve, is in pay group E1-E3, or is non-Hispanic White. As a result, the model with two-way interactions does not achieve statistical balance. In part, this is probably due to the extremely large sample in which small effects test out as statistically significant, and, in part, to misspecification of the model itself. In particular, there may be higher order interactions. Using a regression tree may be one way of finding these.

**Regression Tree**

Using the same four variables as above—service, pay group, gender, and race-ethnicity—we fit a CART model with this code:

t1 <- rpart(resp ~ xcpay1r + xreth4r + xsexr + xsrrcr,

method = "class",

control = rpart.control(minbucket = 250, cp=0),

data = sofr.d1.elig)

The tree with 13 terminal nodes is shown in Figure 16.2. As is apparent from the figure, the structure has some complicated combinations. Table 16.8 gives the descriptions of the nodes. The CART classes are numbered differently by the print method than in the object t1$where. The highest ranking officers have the highest response rates; this is reflected in class 25 (7) containing pay groups O4-O6 which has a response rate of 0.631 (unweighted) and 0.672 (weighted). Enlisted personnel did not respond well—CART put all E1-E6’s in class 2, which has unweighted and weighted rates of 0.272 and 0.321. Among higher paid personnel, Marines are some of the poorest responders. For example, class 21 (104), containing E7-E9, non-Hispanic Whites in the Marine Corps Reserve had an unweighted rate of 0.385 (0.410 weighted). The numbers of persons in the CART classes range from 296 to 41,380, which are obviously far from the more nearly equal-sized classes in the propensity analysis. Note that 41,380 of the 66,148 eligibles (62.6%) are in the same class and are assigned the same response rate.

Since the classes formed by the regression tree seem to capture the complexity of the response process better than the logistic model, the classes in Table 16.8 will be used for nonresponse adjustment. We used weighted response rates to make the weight adjustment in each class. The weighted values are shown in the last column of Table 16.8. As mentioned in Chapter 13, not all practitioners will agree on whether the weighted or unweighted rates should be used. Using the weighted rates is, in a sense, a compromise solution. Conditional on the classes formed, the weighted rates are model-unbiased under a model in which every person in a class has a common probability of responding. They are also approximately unbiased estimates of the population response rates in repeated sampling given the particular set of classes used.

**Table 16.8** Nonresponse adjustment classes created using a regression tree.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CART class (t1$where) | CART class (print) | Description | No. of persons | Unweighted response rate | Weighted response rate |
| 2 | 2 | E1-E3, E4, E5-E6 | 41,380 | 0.272 | 0.321 |
| 7 | 48 | E7-E9, W1-W5, O1-O3, Minority, Marine Corps Reserve | 669 | 0.426 | 0.406 |
| 9 | 98 | E7-E9, W1-W5, O1-O3, Minority, Army National Guard | 1,125 | 0.453 | 0.481 |
| 12 | 396 | E7-E9, W1-W5, O1-O3, Minority, Army Reserve, Female | 473 | 0.469 | 0.471 |
| 14 | 794 | E7-E9, Minority, Army Reserve, Male | 562 | 0.496 | 0.485 |
| 15 | 795 | W1-W5, O1-O3, Minority, Army Reserve, Male | 602 | 0.508 | 0.517 |
| 16 | 199 | E7-E9, W1-W5, O1-O3, Minority, Naval Reserve | 296 | 0.554 | 0.525 |
| 17 | 25 | E7-E9, W1-W5, O1-O3, Minority, Air National Guard, Air Force Reserve | 1,239 | 0.530 | 0.544 |
| 21 | 104 | E7-E9, non-Hispanic White, Marine Corps Reserve | 494 | 0.385 | 0.410 |
| 22 | 105 | W1-W5, O1-O3, non-Hispanic White, Marine Corps Reserve | 561 | 0.510 | 0.509 |
| 23 | 53 | E7-E9, W1-W5, O1-O3, non-Hispanic White, Air National Guard, Air Force Reserve | 3,321 | 0.578 | 0.612 |
| 24 | 27 | E7-E9, W1-W5, O1-O3, non-Hispanic White, Army National Guard, Army Reserve, Naval Reserve | 5,941 | 0.586 | 0.586 |
| 25 | 7 | O4-O6 | 9,485 | 0.631 | 0.672 |



**Figure 16.2** Regression tree to predict response based on the four variables available for respondents and nonrespondents.

**16.7 Calibration to Population Counts**

The final weighting step in this project will be calibration to some of the available population counts. The statistical function that calibration serves here is mainly to reduce standard errors. Since military administrative records should be accurate, there should be no systematic over- or under-coverage to be corrected. In addition, calibration has some cosmetic appeal here. Having estimated counts exactly equal to ones from administrative personnel records will give the survey results face validity—a feature that may be important to many data users. There are two major operational questions that must be addressed:

* Which variables and/or combinations of variables should be used for calibration?
* How should missing values for the calibrating variables be handled in the sample file and the file of population counts?

The code for completing the analyses sketched below is in the file 16.3 Solution calibration adj.R on the website.

Other questions that we will not address here, but would be important in a real survey, are:

* For what time period should population counts be made when there is a delay between sample selection and data collection?
* Which persons should be counted to get the controls?

Administrative record databases are typically updated periodically—once a month, once a quarter, etc. There may also be a lag between the time period of the database and the time at which it is available for tabulation. This means that population counts may not be for the time period when data are collected. In addition, data collection may extend across two or more updates of the administrative data. For example, there might be a 2-month lag between sample selection and data collection, the survey period may last 10 weeks, and the administrative records may be updated once a month. When such a lag occurs, the persons who are surveyed will be the “survivors”, i.e., the ones who were in the frame when the sample was selected and are still eligible when data are collected. No new entrants to the population would be included in the sample. If the population counts are made close to the time of data collection and include all persons who are eligible based on the survey rules, then the counts would include the new entrants who had no chance of being sampled. If we calibrate to these counts, we are saying that the attitudes of the new entrants can be predicted by those of the sample persons who have been in the population longer. Another option would be to tabulate the control counts using only persons who have been in the military for at least two months, if that is the amount of lag between sampling and data collection. In some surveys, like those of the U.S. household population, such selective tabulations may not be feasible.

In this project, we will use the population counts as given in the RCCPDS57.sas7bdat file. As noted in Chapter 12, this file came from matching the sampling frame to the most current personnel file available as of the start of the data collection period. That is, the counts are those of the survivors. Thus, these counts should cover only eligible cases.

**Identifying Variables to Use**

The file of population counts contains combinations of service, gender, pay group, race-ethnicity, education, marital status, and length of activation. All of these are categorical and can be used singly or in any number of combinations. We could, for example, use only the marginal counts of service, pay group, and gender. Or, we could use service × pay group and service × gender; or, service × pay group × gender. Some modeling is a useful approach to guide the decision. The goal will be to determine one set of weights that is reasonably efficient for the important variables measured in the survey. We have six analysis variables listed in Table 16.1 (RA006a, RA006B, RA008, RA115, RA118, and RA112RA) to aid in making the decision.

To do the modeling we created several binary variables. Satisfaction with compensation (RA006a) and type of work (RA006b) were coded as Satisfied/(Very Satisfied) = 1 and 0 otherwise. Likelihood of re-enlisting (RA008) was coded as (Likely/Very likely) = 1 and 0 otherwise. Preparation for job (RA115) was coded as (Well prepared)/(Very well prepared) = 1 and 0 otherwise. Level of stress was coded as (More than usual)/(Much more than usual) = 1 and 0 otherwise. Finally, days in compensated status (RA112RA) was used as a continuous variable.

Rather than fitting binary regressions where the form of the predictors is specified in advance, we again used regression trees to allow the algorithm to identify the more important variables and combinations of levels for prediction. Since the intention to re-enlist is a key variable in this survey, we present those results here. Figure 16.3 shows the regression tree for predicting whether a person is Likely/Very likely to re-enlist. The code for computing the tree and drawing the figure is:

t1 <- rpart(ra008R ~ xsrrcr + xsexr + xcpay1r + xreth4r +

sred + srmarst + xact2r,

method = "class",

control = rpart.control(minbucket = 250, cp=0),

data = datafile)

plot(t1, uniform=TRUE, compress=TRUE, margin = 0.1, branch=0)

text(t1, use.n=TRUE, all=TRUE,

digits=15,

cex=1.2,

pretty=1.2,

fancy=TRUE,

fwidth=0.7,

xpd = TRUE,

font = 3)

The parameter branch=0 gives a tree with V-shaped branches, which, in this case, makes the branch labels easier to read. Pay group, branch of service, whether a person had been activated for 30 days or more, and marital status are included in the tree; gender, race-ethnicity, and education are not.



**Figure 16.3** Regression tree for predicting likelihood of re-enlisting

Table 16.9 Variables included in regression trees for predicting six analysis variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Predictors | | | | | | |
| Name | Analysis variable | Service | Gender | Pay group | Race-ethnicity | Education | Marital status | Activation |
| RA006A | Compensation |  |  |  |  |  |  |  |
| RA006B | Type of work |  |  |  |  |  |  |  |
| RA008R | Re-enlist |  |  |  |  |  |  |  |
| RA115R | Preparation |  |  |  |  |  |  |  |
| RA118 | Stress |  |  |  |  |  |  |  |
| RA112RA | Paid status |  |  |  |  |  |  |  |

Table 16.9 summarizes which variables were included in the trees for predicting the six analysis variables. Gender was selected only to predict whether people felt prepared to do their jobs. Examination of the individual trees shows that service × pay and service × activation interactions are always present. Often there are more complicated interactions, as in Figure 16.3 where there is a combination of service, pay group, activation, and marital status. However, including 3-way and 4-way interactions would lead to samples that are very thin in some combinations of levels even though there are over 25,000 respondents. Based on these results, we decided to use a calibration model with

* main effects for service, gender, pay group, race-ethnicity, education, marital status, and actvation;
* interactions for service × pay and service × activation.

Although gender only appears once in Table 16.9, we include it for the cosmetic benefit of matching the administrative record count for males and females.

**Imputing for Missing Values**

Tables 16.6a and 16.6b showed that the file from which population counts were made had missing values for some persons for service, race-ethnicity, gender, education, marital status, and activation. The percentage of persons with missing values ranged from 0.04% for service to 3.3% for education. To impute for the missing values, we need only impute a covariate value whenever it was missing in the RCCPDS57.sas7bdat file. For example, here is one record in the file that had a missing value for service:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| service | gender | pg\_group | raceth | educcat | marit | activatd | count |
| . | 2 | 2 | 2 | 4 | 5 | 3 | 159 |

To impute the missing service, a random draw is made from allowable the codes in proportion to the population code-counts for the non-missing records. The R code for doing the population count imputations is in the function, impute, in the file 16.3 Solution calibration adj.R.

The sample file also has some records with missing data on the covariates that will be used for calibration. Table 16.6b shows that 2.1% of the 25,559 respondents are missing education, 0.1% are missing marital status, and 0.5% are missing the activation field. Any missing value for a sample respondent was imputed with a random draw from the allowable codes for a variable. The draws were made in proportion to the distribution among the codes for persons with non-missing data. The R code for doing the sample imputations is in the function, impute.sam, which is also in 16.3 Solution calibration adj.R.

These imputation methods are straightforward and could be criticized as not accounting for any multivariate relationships among different variables. Given the small amount of missing data for all variables, we elected to keep the methods simple.

**GREG Estimation**

Using the files of sample respondents and population counts with all missing values imputed, we calibrated to the population totals using a GREG estimator. When using the calibrate function in R survey, some care is needed to be sure that the vector of population totals is in exactly the same order as is being used internally by calibrate. The function model.matrix will create the model-matrix of covariates that calibrate uses for a particular formula. In this application, we check the order with:

# check how design matrix is formed in calibrate

mm <- model.matrix(~ as.factor(xsrrcr) \* as.factor(xcpay1r)

+ as.factor(xsrrcr) \* as.factor(xact2r)

+ as.factor(sred)

+ as.factor(xsexr)

+ as.factor(xreth4r)

+ as.factor(srmarst),

data = sofr.cal)

dimnames(mm)[[2]]

The last statement lists the column names of the model-matrix. The interactions are in “row-major” order. For example, the first five values of the service  paygrade interaction are

"as.factor(xsrrcr)2:as.factor(xcpay1r)2"

"as.factor(xsrrcr)3:as.factor(xcpay1r)2"

"as.factor(xsrrcr)4:as.factor(xcpay1r)2"

"as.factor(xsrrcr)5:as.factor(xcpay1r)2"

"as.factor(xsrrcr)6:as.factor(xcpay1r)2"

That is, service is incremented before paygrade. The code for putting the population controls in the correct order and for computing the GREG weights follows. Prior to this code counts for service  paygrade and service  activation were made and stored in the objects svc.pay1 and svc.act1.

# reorder the pop totals for the interaction terms to match

# way that calibrate creates model matrix

svc.pay1 <- svc.pay[order(svc.pay[,2]),]

svc.act1 <- svc.act[order(svc.act[,2]),]

del1 <- svc.pay1[,1]==1 | svc.pay1[,2]==1

del2 <- svc.act1[,1]==1 | svc.act1[,2]==1

pop.tots <- c(N,

svc[-1,2],

pay[-1,2],

activated[-1,2],

educ[-1,2],

gender[-1,2],

raceth[-1,2],

marital[-1,2],

svc.pay1[!del1,3],

svc.act1[!del2,3])

sam.lin.ub <- calibrate(design = sofr.cal.dsgn,

formula = ~ as.factor(xsrrcr) \* as.factor(xcpay1r)

+ as.factor(xsrrcr) \* as.factor(xact2r)

+ as.factor(sred)

+ as.factor(xsexr)

+ as.factor(xreth4r)

+ as.factor(srmarst),

population = pop.tots,

bounds = c(-Inf,Inf),

calfun = c("linear")

)

Table 16.10 gives some summary statistics on the weights after each step in the process. The mean weight is about the same before and after the GREG step, while the range is larger for the GREG weights than for the nonresponse-adjusted weights. The sum of the weights is smallest after the GREG step (801,809) accounting for the fact that some persons became ineligible between sampling and data collection and that the control totals are for the survivors only.

Table 16.10 Summary of weights and counts of persons after each step.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weighting step | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Sum | Persons |
| Base | 1 | 2.201 | 5.049 | 12.14 | 14.27 | 178.3 | 870,373 | 71,701 |
| Adjusted for unknown eligibility | 1.023 | 2.251 | 5.05 | 12.3 | 14.59 | 182.3 | 813,342 | 66,148 |
| Adjusted for nonresponse | 1.521 | 4.746 | 14.63 | 31.82 | 34.31 | 514.7 | 813,342 | 25,559 |
| GREG | 1.199 | 4.672 | 13.43 | 31.37 | 30.91 | 613.4 | 801,809 | 25,559 |

In this solution weight trimming was not used, although some practitioners might consider it. Although the range of final weights is fairly large—1.199 to 613.4—the base weights began with a wide range owing to the highly differential sampling rates that were used. The base weights were adjusted to reflect substantially different response rates among some types of personnel. Consequently, the final weights have a wide range. This is necessary to correct nonresponse bias for some subgroups. However, the 99th quantile of the final weights is about 385 while the maximum final weight is 613.4. Trimming of the largest 1% of weights might reduce SE’s for full population estimates without introducing too much bias, but estimates for the subgroups with very low response rates might then be biased. As usual, we are faced with conflicting goals with no unique way of achieving them.

**16.8 Writing Output Files**

The resulting file with the GREG weights can be written to comma delimited (csv) text files for use in other statistical software. The code below appends the GREG weight to the file, selects fields for output, and writes the text files. The write.foreign function in the foreign package will also write code to be used in importing the csv files into a few other packages. We illustrate the process below for SAS and Stata.

# append GREG weights to data file of 25,559 respondents

sofr.cal$d3 <- weights(sam.lin.ub)

# specify fields for the text, SAS, and Stata files

fields <- c("rec.id", "nr.class", "respstat", "stratum",

"nsamp", "nstrat", "v.strat",

"srmarst", "sred", "xsrrcr", "xact2r",

"xreth4r", "xsexr", "xcpay1r",

"ra006a", "ra006b", "ra008", "ra115", "ra118", "ra112ra",

"pred.logit", "p.class.10", "unwt.rr", "wt.rr",

"d0", "d1", "a1",

"d2", "a2",

"d3")

write.foreign(df = sofr.cal[, fields],

datafile = paste(file\_loc2, "sofr.cal.sas.csv", sep=""),

codefile = paste(file\_loc2, "sofr.sas", sep=""),

package = "SAS")

write.foreign(df = sofr.cal[, fields],

datafile = paste(file\_loc2, "sofr.cal.stata.csv", sep=""),

codefile = paste(file\_loc2, "sofr.ado", sep=""),

package = "Stata")

The reader can consult the programs, 16.1 Solution bwt-unknown adj.R, 16.2 Solution NR adj.R, and 16.3 Solution calibration adj.R, to see how the different variables were created.

Although the data can be imported into statistical packages other than R, a worry is that the other packages do not have built-in procedures that recognize that the weights were computed via the GREG procedure. This, typically, means that linearization variance estimates will be computed using the ultimate cluster method discussed in Chapter 15 that does not use the correct set of residuals. As a result, linearization SE’s computed from the other packages will not generally be correct. This problem can be avoided if replication is used. In that case, the set of replication weights can appropriately reflect the different steps in weighting, particularly the type of calibration that was used. The replicate weights are included with the data set, and a package like SAS or Stata needs only to be told which method of replication was used—jackknife, BRR, or the bootstrap—in order to produce legitimate SE’s.

**16.9 Example Tabulations**

Finally, in this section we present a few simple tabulations using the file with the final weights. R Code is in the file, 16.4 Example tabulations.R. The proportions of persons responding in the categories of the re-enlistment question (ra008) can be estimated with:

# proportions for re-enlistment item

reenlist <- svymean(~ as.factor(ra008), design = sam.lin.ub,

na.rm = TRUE)

# format with row labels

print(ftable(reenlist,

rownames = list(c("Very unlikely",

"Unlikely",

"Neither likely nor unlikely",

"Likely",

"Very likely")

)

), digits = 3)

The function ftable allows labels to be used for the printed output. The results, even with the use of ftable which improves the appearance, is not beautiful:

Very unlikely mean 0.06608

SE 0.00357

Unlikely mean 0.10924

SE 0.00428

Neither likely nor unlikely mean 0.09280

SE 0.00386

Likely mean 0.31757

SE 0.00611

Very likely mean 0.41431

SE 0.00619

If a table is needed for a report, one option is to import the output into a spreadsheet where more attractive formatted can be applied. Suppose that the result of print(ftable(reenlist, …)) is saved to an object called out. Code that will convert out to a data frame, put the proportion and SE side by side, and write the result to a file called table.csv is:

out <- data.frame(out)

out <- cbind(out[1:5,], out[6:10,])

out <- out[, c(1,3,6)]

dimnames(out)[[2]] <- c("Response", "Proportion", "SE")

write.csv(out, file = "c:\\table.csv")