Adjusting sample weights for linkage-eligibility using SUDAAN

Dean H. Judson[1], Jennifer D. Parker[1], and Michael D. Larsen[2]

[1]National Center for Health Statistics, Special Projects Branch, Office of Analysis and Epidemiology

[2]The George Washington University, Department of Statistics

Suggested citation: Judson DH, Parker JD, Larsen MD. Adjusting sample weights for linkage-eligibility using SUDAAN. National Center for Health Statistics, Hyattsville Maryland. May 2013. Available at the following address: http://www.cdc.gov/nchs/data/datalinkage/adjusting_sample_weights_for_linkage_eligibility_using_sudaan.pdf

Introduction

The National Center for Health Statistics (NCHS) record linkage program increases the scientific value of the Center's population-based surveys. Recent record-linked file data releases include linkage of NCHS surveys to the National Death Index (NDI) with mortality follow-up, to enrollment and claims data for Medicare and Medicaid programs from the Centers for Medicare and Medicaid Services (CMS), and to enrollment and claims data from the Social Security Administration (SSA). The linked mortality files have been used to examine mortality and cause of death for many factors identified on NCHS surveys. Detailed descriptions of the methodologies used to create the NCHS linked data files are available elsewhere (1).

Although many NCHS surveys have been linked to administrative records, not all survey participants are linkage-eligible. Further, linkage-eligibility can differ for each survey and administrative data linkage and over time. In general, if a survey respondent did not provide sufficient personally identifying information (PII) or explicitly refused to be linked he or she is not linkage-eligible. In particular, for many linkages, the provision of a Social Security Number (SSN) is a criterion for linkage-eligibility. Miller and colleagues examined the proportion of National Health Interview Survey (NHIS) respondents refusing to provide SSN from 1997 to 2009. They report that the proportion of respondents refusing generally increased until 2006 and began to decrease after 2007, when the survey began collecting the last four digits of the SSN instead of the full nine digits (2). Linkage-eligibility is distinct

from program-eligibility. A survey respondent can be linkage-eligible but not match to, say, Medicare claims records because he or she is not in the Medicare program. A survey respondent who is not in the Medicare program is typically not considered a non-responder in analyses of Medicare-linked data files. A small number of respondents who are program-eligible are not successfully matched to the administrative data – possibly because of errors in the necessary PII. These survey respondents can be considered, along with those who are not linkage-eligible, as non-responders in the linked data files. Documentation on linkage-eligibility and match-rates are provided for each NCHS linkage (1). For the purposes of investigating methods for using the NCHS linked data, the refusal by a survey participant to provide sufficient PII for linkage or to allow linkage of their survey data is considered "non-response".

For its population health surveys, NCHS creates sample weights based on probabilities of selection into the survey, with adjustments for non-response and Census post-stratification. Post-stratification involves adjusting the sample weights within groups to match totals in the population, such as Census population totals or population estimates. NCHS public use sample weights are computed for all survey responders and users of the NCHS surveys are encouraged to use these sample weights in their calculations. (Indeed, it is important that the weights be used in order to produce estimates that accurately represent the target population.)

However, only the linkage-eligible survey respondents can be used when analyzing NCHS linked data, not the full sample. Survey respondents who provide sufficient PII for linkage are not a random sample of respondents. Instead, the linkage-eligible are a self-selected subset of the initial survey respondents. Bias is caused when the non-respondents differ systematically from the respondents in terms of some characteristics of interest. Bias can occur in estimates of totals, means, proportions, coefficients in regression-type analyses (3). Both by increasing uncertainty and by causing bias, non-response can invalidate or weaken conclusions based on survey data.

Of critical importance is whether non-response rates (i.e., non-linkage to administrative data) vary by subsets of the sample and whether responses to outcomes of interest vary across these groups (4). If so, there is potential for non-response bias. Non-response weighting adjustments can be effective if the non-response weighting is related to the variables that influence the probability of non-response and to the variables being studied.

For developing and testing reweighting models, we have found PROC WTADJUST in SUDAAN (5) to be useful. This SUDAAN procedure is designed to work with complex surveys and is flexible enough for implementing different approaches to re-weighting. In addition, as shown below, multiple approaches can be easily applied and compared. NCHS uses a SAS-callable version of SUDAAN, but a stand-alone version is also available.

This document is organized as follows. We begin with a brief overview of SUDAAN's PROC WTADJUST. Second, we describe the data files used in the examples. Third, we present 3 successful examples using PROC WTADJUST (we will also illustrate a model that does not converge, for contrast). Finally, we provide some suggestions for examining the adjusted sample weights, to enhance confidence that the weight adjustment model chosen will lead to valid inferences.

We assume that the reader is familiar with standard SAS program language conventions, e.g., libnames and filenames, in the examples that follow. We advise readers to read the SUDAAN documentation for a detailed understanding of the WTADJUST procedure and its implementation. Although we show 3 examples of its use below, the best approach for a particular project may require different applications of the procedure. Furthermore, other software can be used to re-weight survey data (6) and other methods can be used to address the issue of linkage-eligibility (7).

PROC WTADJUST

SUDAAN is a statistical package designed to correctly handle data analysis for data from complex sample designs. Within SUDAAN, PROC WTADJUST is a module designed specifically for nonresponse and post-stratification adjustments. Weight adjustments are created using a model-based, calibration approach. There are two ways to use PROC WTADJUST: 1) to correct for non-response using a model-based approach ("nonresponse" option) and 2) to directly post-stratify to external control totals ("post-stratify" option). Control totals are population estimates or counts calculated for specific cross-categories of subgroups, often defined by race and ethnicity, age, and sex categories. If the variables that were used to define control totals for the original sample weights are used in PROC WTADJUST, then applying the

nonresponse option and fitting a model using all cross-categories of the variables defining the control totals will produce the same adjusted sample weights as using the post-stratify option with those control totals. We illustrate both options with specific examples.

Data files used in the examples

Feasibility files.

Publicly available feasibility data files for the linkages of NCHS survey data to the Centers for Medicare and Medicaid Services (CMS) Medicare claims can be downloaded directly from the NCHS website. The NCHS-CMS-Medicare feasibility files provide a limited set of variables that can be used to determine the maximum available sample size for each linked file. They also can be used to assess the potential impact of linkage-ineligible records. These files are especially useful to researchers considering whether to initiate a Research Data Center (RDC) proposal to analyze the restricted-use linked NCHS-CMS-Medicare files. Each feasibility file is NCHS survey and survey year specific.

The following information is included on each feasibility study file:

- NCHS public use data file identifier (for the National Health Interview Survey (NHIS), this is called `PUBLICID`);
- Survey respondent eligibility and final match status (`CMS_MEDICARE_MATCH`**, Figure 1**);
- Variables specifying on which CMS Medicare data files the successfully linked survey respondent has information.

The feasibility study files do not contain any specific information about CMS Medicare benefits. Data users will need to use information from the public use survey or external information to approximate the number of respondents with a specific condition. For example, by merging with the public use 2005 sample adult file, one could identify how many of the respondents who reported that they had been told they have hypertension (`HYPEV`) are also linkage eligible and have data from one or more Medicare files.

Figure 1. Variable excerpt from NCHS-CMS Medicare Feasibility file documentation.

CMS_MEDICARE_MATCH CMS MEDICARE MATCH STATUS

Type: Numeric **Width:** 1 **File Position:** 15

Possible values:

1 Linkage-eligible & linked to one or more years of CMS Medicare administrative data

2 Linkage-eligible & not linked to any year of the CMS Medicare administrative data

3 Linkage-eligible & linked to one or more years of the CMS Medicare administrative data

- Child survey participant (less than 18 years of age at time of survey) & turned 18 during the administrative data coverage period - CMS Medicare administrative data available prior to the participant's 18th birthday, but information indicating whether child linked after 18th birthday is not available

9 Ineligible for CMS Linkage

Usage Notes:

Survey respondents are ineligible for linking to CMS administrative records if they are missing key identification data and/or if they refused to provide their Social Security or Medicare Health Insurance Claim number at the time of the survey interview or did not have a Social Security number verified by the Social Security Administration's Enumeration Verification System.

In accordance with NCHS Ethics Review Board (ERB) guidelines, NCHS will no longer release linked administrative data for child survey participants if the administrative data was generated for program participation, claims or other events occurring on or after their 18th birthday.

SOURCE: NCHS-CMS Medicare Feasibility Study File Codebook (Updated 12/2012)
http://www.cdc.gov/nchs/data/datalinkage/cms_medicare_feasibility_data_codebook.pdf

The NCHS-CMS-Medicare feasibility file can be found at: http://www.cdc.gov/nchs/data_access/data_linkage/cms/cms_medicare_feasibility.htm. For our examples we use the feasibility file for the 2005 NHIS. We put it into the SAS library NHIS05, and refer to it as NHIS05.FEAS. Starting in 2013, an additional category was added, "3 LINKAGE-ELIGIBLE – CHILD SURVEY PARTICIPANT", which flagged persons who, upon turning age 18, would become ineligible. For this exercise, we treat them as ineligible.

2005 National Health Interview Survey, Public-use Person file.

The public use 2005 NHIS person file can be found at: http://www.cdc.gov/nchs/nhis/nhis_2005_data_release.htm.

From the 2005 NHIS person file, the following public use variables are used: PUBLICID, AGE_P (respondent's age at survey interview), HISCODI2 (race and Hispanic origin, which is renamed "RACEETH" to be more descriptive), SEX, EDUC1 (education), PHSTAT (which is recoded into a logical variable FAIRPOORHEALTH for assessment), REGION (West, Northeast, South, Midwest), WTFA (the original design-based sample weight), STRATUM, and PSU (primary sampling unit).

In our examples, we use the variable PUBLICID, which is on both the feasibility file and the 2005 NHIS person file. Because public-use NHIS files have had varying variable names in different years, PUBLICID may need to be created in a year-specific way. See the following for more information: (http://www.cdc.gov/nchs/data/datalinkage/important_information_on_merging_nchs_restricted_and_public_use_survey_data.pdf).

To create the data file, sort and merge the files and code the variables. For our examples, the input data files and created data files were kept in the SAS library NHIS05. The public use 2005 NHIS person file is called NHIS05.PERSONSX.

```
/* DEFINE FORMATS FOR LATER USE */
      PROC FORMAT ;
```

```
VALUE EDUCF
1=”< High school”
2=”High school or GED or some college”
3=”College degree” ;

VALUE AGECATF
1=”UNDER 18 YEARS”
2=”18-44 YEARS”
3=”45-64 YEARS”
4=”65 YEARS OR OLDER” ;

VALUE REGIONF
1=”NORTHEAST”
2=”MIDWEST”
3=”SOUTH”
4=”WEST” ;

VALUE RACEETHF
1="HISPANIC"
2="NON-HISPANIC WHITE"
3="NON-HISPANIC BLACK"
4="ALL OTHER RACES" ;

VALUE FAIRPOORF
1="FAIR OR POOR HEALTH"
```

```
            0="GOOD, VERY GOOD, EXCELLENT HEALTH"
            .="MISSING" ;
/* STEP 1:  MERGE THE FEASIBILITY AND PUBLIC USE DATA FILES.*/

PROC SORT DATA=NHIS05.FEAS ;     BY PUBLICID ;
PROC SORT DATA=NHIS05.PERSONSX ; BY PUBLICID ;

DATA NHIS05.MERGED NOTMERGED1 NOTMERGED2 ;
     MERGE
          NHIS05.FEAS (IN=A KEEP=PUBLICID CMS_MEDICARE_MATCH)
          NHIS05.PERSONSX (IN=B KEEP=PUBLICID AGE_P HISCODI2 PHSTAT SEX
          REGION EDUC1 WTFA STRATUM PSU) ;
     BY PUBLICID ;
     IF A=1 AND B=1 THEN OUTPUT NHIS05.MERGED ;
     IF A=0 AND B=1 THEN OUTPUT NOTMERGED1 ;
     IF A=1 AND B=0 THEN OUTPUT NOTMERGED2 ;

/* CHECK THE LOG FILE.  NO RECORDS SHOULD BE OUTPUT TO NOTMERGED1 OR
NOTMERGED2. */

/* STEP 2: DEFINE THE LINKAGE-ELIGIBILITY (MISSING DATA) INDICATOR,
LINKABLE.  RECODE THE VARIABLES.  CREATE VARIABLE FORMATS. */

DATA NHIS05.MERGED ; SET NHIS05.MERGED ;
     LINKABLE=CMS_MEDICARE_MATCH IN (1,2) ;  /*DEFINES LINKAGE-ELIGIBILTY
     */
     FAIRPOORHEALTH=(PHSTAT IN (4,5)) ;
     IF PHSTAT=7 OR PHSTAT=8 OR PHSTAT=9 THEN FAIRPOORHEALTH=. ;
```

```
      /* DEFINE THE FAIR OR POOR HEALTH VARIABLE */ ;

      AGE_CAT=. ;
      IF 0 LE AGE_P LE 17 THEN AGE_CAT=1 ;
      IF 18 LE AGE_P LE 44 THEN AGE_CAT=2 ;
      IF 45 LE AGE_P LE 64 THEN AGE_CAT=3 ;
      IF AGE_P > 64 THEN AGE_CAT=4 ;

      AGE_P2 = AGE_P*AGE_P ;  * AGE-SQUARED ;

      EDUC=. ;
      IF EDUC1 IN (0,1,2,3,4,5,6,7,8,9,10,11) THEN EDUC=1 ; *<HIGH SCHOOL
      ;
      ELSE IF EDUC1 IN (12,13,14,15,16,17) THEN EDUC=2 ; *HIGH SCHOOL OR
      GED OR SOME COLLEGE ;
      ELSE IF EDUC1 IN (18,19,20,21) THEN EDUC=3 ; *COLLEGE DEGREE ;

      /* CONVERT PUBLICID TO NUMERIC FOR USE IN SUDAAN */
      ID=PUBLICID*1 ;

      RACEETH=HISCODI2; * CREATES A MORE DESCRIPTIVE VARIABLE NAME ;

FORMAT AGE_CAT AGECATF. RACEETH RACEETHF. SEX SEXF. REGION REGIONF. EDUC
EDUCF. FAIRPOORHEALTH FAIRPOORF. ;
```

Examples

Example 1: MARGINAL MODEL

The marginal, or main effects, model is fit using AGE_CAT, HISCODI2 (renamed RACEETH), SEX, REGION and EDUC. By definition of a marginal model, in this model no interaction terms are specified. The factors REGION and EDUC (education, in three categories) are often related to the propensity to agree to linkage as well as the underlying survey design, where strata are formed based on geographic and socio-economic characteristics of locations within the U.S.

After fitting the model, using the output statement, the adjusted weights produced by the model are saved in a new temporary SAS dataset called MATCH1. Next, MATCH1 is merged to the original files, NHIS05.MERGED, so the adjusted weights can be evaluated and used in an analysis.

```
/* MARGINAL MODEL:  THIS MODEL INCLUDES MAIN EFFECTS ONLY*/

PROC WTADJUST DATA=NHIS05.MERGED DESIGN=WR ADJUST=NONRESPONSE NOTSORTED ;
     NEST STRATUM PSU ;
     WEIGHT WTFA ;
     CLASS AGE_CAT HISCODI2 SEX REGION EDUC / INCLUDE=MISSING ;
     REFLEVEL AGE_CAT=2 RACEETH=2 SEX=1 REGION=1 EDUC=2 ;
     MODEL LINKABLE=AGE_CAT RACEETH SEX REGION EDUC ;
     IDVAR LINKABLE AGE_CAT RACEETH SEX REGION EDUC ID ;
     PRINT BETA SEBETA P_BETA MARGADJ / BETAFMT=F10.4 SEBETAFMT=F10.4 ;
     OUTPUT /PREDICTED=ALL FILENAME=MATCH1 FILETYPE=SAS REPLACE ;
RUN ;

PROC SORT DATA=MATCH1 ; BY ID ;
RUN ;
```

```
PROC SORT DATA=MERGED ; BY ID ;
RUN ;

/* MERGE THE WEIGHT FILE TO THE MERGED FILE.  RENAME ADJFACTOR AND
WTFINAL, WHICH ARE SUDAAN INTERNAL VARIABLES, TO ADJFACT_MARGINAL AND
WTFIN_MARGINAL, SO THAT WE CAN COMPARE THE ADJUSTMENTS.*/

DATA NHIS05.MERGED_WEIGHTED ;
     MERGE NHIS05.MERGED MATCH1 (KEEP=ID ADJFACTOR WTFINAL) ;
     BY ID ;
     ADJFACT_MARGINAL=ADJFACTOR ;
     WTFIN_MARGINAL=WTFINAL ;
RUN ;
```

Example 2: FULLY SATURATED MODEL.

Post-stratifying to external population control totals is another way to adjust sample weights. Although PROC WTADJUST has this ability, this adjustment can also be obtained using a fully saturated model if the variables and their categories are the same as those that would be used for post-stratification. This is because the sum of the original weights within each cell is the population total for that cell. Models like this are typically referred to as "saturated" models and are fitted by including all possible interactions among the variables. However, specifying too may variables may not work, in part because the interactions will lead to cells with too few observations for stable estimates. We demonstrate this with a model that fails to converge, and then we show two alternatives that use fewer interaction terms.

```
/* FULLY SATURATED MODEL:  THIS MODEL INCLUDES ALL COMBINATIONS OF ALL
VARIABLES AND FAILS TO CONVERGE */

PROC WTADJUST DATA=NHIS05.MERGED_WEIGHTED DESIGN=WR ADJUST=NONRESPONSE
NOTSORTED ;
```

```
      NEST STRATUM PSU ;
      WEIGHT WTFA ;
      CLASS AGE_CAT RACEETH SEX REGION EDUC / INCLUDE=MISSING ;
      REFLEVEL AGE_CAT=2 RACEETH=2 SEX=1 REGION=1 EDUC=2 ;
      MODEL LINKABLE=AGE_CAT*RACEETH*SEX*REGION*EDUC ;
      IDVAR LINKABLE AGE_CAT RACEETH SEX REGION EDUC ID ;
      PRINT BETA SEBETA P_BETA MARGADJ / BETAFMT=F10.4 SEBETAFMT=F10.4 ;
      OUTPUT /PREDICTED=ALL FILENAME=MATCH2 FILETYPE=SAS REPLACE ;
RUN ;
```

As indicated above, as the iteration proceeds to fit this model, some of the cells are empty, and as a result, the model coefficients in the underlying logistic regression approach infinity. Estimates that are very large (typically with very large standard errors) can occur when every person in a cell is a respondent. Estimates that are very negative can occur when every person in a cell is a nonrespondent. While very fine adjustment cells are generally desirable, these adjustments are too fine.

Example 3: SATURATED MODEL

This example is a compromise between the above two. Now that we know that a model with interactions among all variables will not converge, we fit a model with all possible interaction terms for categories of age, race and sex (known as a "saturated" model) and still include the variables `REGION` and `EDUC` independently.

```
/* SATURATED MODEL: THIS MODEL INCLUDES ALL POSSIBLE INTERACTIONS OF AGE,
RACE/ETHNICITY, AND SEX AND ONLY MAIN EFFECTS FOR REGION AND EDUC */

PROC WTADJUST DATA=NHIS05.MERGED_WEIGHTED DESIGN=WR ADJUST=NONRESPONSE
NOTSORTED ;
      NEST STRATUM PSU ;
```

```
      WEIGHT WTFA ;
      CLASS AGE_CAT RACEETH SEX REGION EDUC / INCLUDE=MISSING ;
      REFLEVEL AGE_CAT=2 RACEETH=2 SEX=1 REGION=1 EDUC=2 ;
      MODEL LINKABLE=AGE_CAT*RACEETH*SEX REGION EDUC ;
      IDVAR LINKABLE AGE_CAT RACEETH SEX REGION EDUC ID ;
      PRINT BETA SEBETA P_BETA MARGADJ / BETAFMT=F10.4 SEBETAFMT=F10.4 ;
      OUTPUT /PREDICTED=ALL FILENAME=MATCH2 FILETYPE=SAS REPLACE ;
RUN ;

PROC SORT DATA=MATCH2 ; BY ID ;
RUN ;

/*NOW MERGE THE WEIGHT FILE BACK ONTO THE MERGED FILE*/
DATA NHIS05.MERGED_WEIGHTED ;
      MERGE NHIS05.MERGED_WEIGHTED MATCH2 (KEEP=ID WTFINAL ADJFACTOR) ;
      BY ID ;
      ADJFACT_SATURATED=ADJFACTOR ;
      WTFIN_SATURATED=WTFINAL ;
RUN ;
```

Example 4: CONTINUOUS AGE MODEL

In our final example, we use continuous covariates AGE_P and AGE_P2, indicating age and age-squared, respectively, while including all possible interactions for the other categorical variables.

```
/* CONTINUOUS AGE MODEL:  THIS MODEL INCLUDES AGE AND AGE-SQUARED AS
CONTINUOUS VARIABLES AND ALL POSSIBLE INTERACTIONS FOR OTHER VARIABLES */
PROC WTADJUST DATA=NHIS05.MERGED_WEIGHTED DESIGN=WR ADJUST=NONRESPONSE
NOTSORTED ;
      NEST STRATUM PSU ;
```

```
      WEIGHT WTFA ;
      CLASS RACEETH SEX REGION EDUC/ INCLUDE=MISSING ;
      REFLEVEL RACEETH=2 SEX=1 REGION=1 EDUC=2 ;
      MODEL LINKABLE=AGE_P AGE_P2 RACEETH*SEX*REGION*EDUC ;
      IDVAR LINKABLE AGE_P AGE_P2 RACEETH SEX REGION EDUC ID ;
      PRINT BETA SEBETA P_BETA NTRIMMED  MARGADJ/ BETAFMT=F10.4
      SEBETAFMT=F10.4 ;
      OUTPUT /PREDICTED=ALL FILENAME=MATCH3 FILETYPE=SAS REPLACE ;
RUN ;

PROC SORT DATA=MATCH3 ; BY ID ;
RUN ;

/* NOW MERGE THE WEIGHT FILE BACK ONTO THE MERGED FILE. AS ABOVE, RENAME
ADJFACTOR AND WTFINAL, WHICH ARE SUDAAN INTERNAL VARIABLES, TO
ADJFACT_CONT_AGE AND WTFIN_CONT_AGE. */
DATA NHIS05.MERGED_WEIGHTED ;
      MERGE NHIS05.MERGED_WEIGHTED MATCH3 (KEEP=ID WTFINAL ADJFACTOR) ;
      BY ID ;
      ADJFACT_CONT_AGE=ADJFACTOR ;
      WTFIN_CONT_AGE=WTFINAL ;
RUN ;
```

Assessment of adjusted weights

An examination of the adjusted weights should be done before their use in analysis. Some approaches are illustrated here, although there are others. First, the basic SUDAAN model results from Example 1 (Marginal model) are shown in **Figure 2**.

Figure 2. Basic SUDAAN output from the marginal model, Example 1

Date: 07-09-2012 SUDAAN Page: 1

Time: 11:38:29 Table: 1

Variance Estimation Method: Taylor Series (WR)
Response variable LINKABLE: LINKABLE
Nonresponse Adjustment
by: Independent Variables and Effects.

Independent Variables and Effects		Beta Coeff.	SE Beta	P-value T-Test B=0	Marginal Weight Adjustment
Intercept		**-0.1693**	**0.0365**	**0.0000**	**2.1289**
AGE_CAT	under 18 years	-0.6819	0.0381	0.0000	1.9661
	18-44 years	0.0000	0.0000	.	2.1607
	45-64 years	0.0563	0.0219	0.0106	2.1891
	65 years or older	0.1389	0.0350	0.0001	2.2897
RACEETH	Hispanic	0.5940	0.0391	0.0000	2.7480
	Non-Hispanic White	0.0000	0.0000	.	1.9927
	Non-Hispanic Black	0.2767	0.0386	0.0000	2.1864
	Non-Hispanic Other	0.4917	0.0617	0.0000	2.6872
Sex	male	0.0000	0.0000	.	2.0730
	female	0.1169	0.0152	0.0000	2.1854
Region	Northeast	0.0000	0.0000	.	2.1764
	Midwest	-0.2345	0.0429	0.0000	1.9180
	South	-0.2183	0.0423	0.0000	2.0258
	West	0.2873	0.0479	0.0000	2.6268
EDUC	.	1.0530	0.0595	0.0000	2.5872
	< high school	0.2888	0.0319	0.0000	2.0303
	high school or GED or some college	0.0000	0.0000	.	2.0715
	college degree	0.1539	0.0279	0.0000	2.2230

The default under SAS/SUDAAN is to define the last category as the baseline category and set the coefficient on baseline category to zero. The other coefficients are interpreted as deviations from the baseline category. The baseline can be changed from its default using the `REFLEVEL` statement, as we have illustrated above and subsequently (5). The `/INCLUDE=MISSING` option in the model code instructs SUDAAN to treat missing values as a separate class, and thus `EDUC=MISSING` gets its own parameter estimate. Although this approach is not generally recommended for statistical analysis, applying other methods (e.g., multiple imputation) or identifying the best way to include non-response in the models was beyond the scope of this project.

WTADJUST provides a marginal weight adjustment column, which indicates the average adjustment to the initial weights for records in each category. Users should examine these values see if any one category has undue influence. As about 50% of respondents were ineligible for linkage in 2005, in general we would expect that these weight adjustments would be around two. Large differences in the marginal weight adjustment across groups could be associated with the creation of large weights and high variability in the distribution of weights, which can be associated with undesirably large standard errors for some estimate and domain combinations.

After inspecting the results provided by SUDAAN additional examinations can be performed. For example, the adjustment cell sizes should be examined; a cell size less than 30 is not generally recommended (4). In addition, correlations and scatter plots of adjusted and unadjusted weights should be examined, to qualitatively identify outliers or other quirks in the adjustment process.

When assessing the results from the WTADJUST procedure, records not eligible for record linkage will have a positive `WTFA` and a zero for `WTFIN_MARGINAL, WTFIN_SATURATED`, and `WTFIN_CONT_AGE`, as well as `ADJFACT_MARGINAL, ADJFACTOR_SATURATED` and `ADJFACT_CONT_AGE`. These zeroes for the adjusted weights will distort summary statistics, correlations and graphics, so the zeroes should be removed before performing these steps. The following code implements some of these checks:

Figure 3. Summary of weight distributions for original weights and for marginal, saturated, and continuous age model weights (SAS output).

Variable	Label	N	Sum	Minimum	Maximum	Variance	Skewness	Kurtosis
WTFA	Weight -FA	98649	291143602	0	19434.00	1664152.47	1.6185274	9.9043452
wtfin_marginal		44650	291143602	1158.37	46481.63	7460528.43	2.2087014	14.2250202
wtfin_saturated		44650	291143602	1136.95	48140.60	7563762.51	2.1796876	13.8933762
wtfin_cont_age		44650	291143602	1131.46	51772.00	7680229.90	2.1965265	13.5749493
adjfact_marginal		44650	97857.42	1.2784220	5.5004675	0.2279472	1.2170257	2.2459365
adjfact_saturated		44650	97810.90	1.3011513	5.5509506	0.2400788	1.2441005	2.1488036
adjfact_cont_age		44650	98177.30	1.3120342	6.8079465	0.3027268	1.5640493	4.1725009

Figure 4. Pair-wise correlations among weights (SAS output, PROC CORR).

Any zero weight has been converted to missing

The CORR Procedure

4 Variables:	WTFA wtfin_marginal wtfin_saturated wtfin_cont_age

Pearson Correlation Coefficients, N = 44650
Prob > |r| under H0: Rho=0

	WTFA	**wtfin_marginal**	**wtfin_saturated**	**wtfin_cont_age**
WTFA Weight - Final Annual	1.00000	0.87816 <.0001	0.87018 <.0001	0.84258 <.0001
wtfin_marginal	0.87816 <.0001	1.00000	0.99221 <.0001	0.96174 <.0001
wtfin_saturated	0.87018 <.0001	0.99221 <.0001	1.00000	0.95756 <.0001
wtfin_cont_age	0.84258 <.0001	0.96174 <.0001	0.95756 <.0001	1.00000

3 Variables:	adjfact_marginal adjfact_saturated adjfact_cont_age

Pearson Correlation Coefficients, N = 44650
Prob > |r| under H0: Rho=0

	adjfact_marginal	**adjfact_saturated**	**adjfact_cont_age**
adjfact_marginal	1.00000	0.96211 <.0001	0.88760 <.0001
adjfact_saturated	0.96211 <.0001	1.00000	0.85933 <.0001
adjfact_cont_age	0.88760 <.0001	0.85933 <.0001	1.00000

```
DATA NHIS05.MERGED_WEIGHTED ;
SET NHIS05.MERGED_WEIGHTED ;
      IF WTFIN_MARGINAL=0 THEN WTFIN_MARGINAL=. ;
      IF WTFIN_SATURATED=0 THEN WTFIN_SATURATED=. ;
      IF WTFIN_CONT_AGE=0 THEN WTFIN_CONT_AGE=. ;
      IF ADJFACT_MARGINAL=0 THEN ADJFACT_MARGINAL=. ;
      IF ADJFACT_SATURATED=0 THEN ADJFACT_SATURATED=. ;
      IF ADJFACT_CONT_AGE=0 THEN ADJFACT_CONT_AGE=. ;
RUN ;
TITLE "ANY ZERO WEIGHT HAS BEEN CONVERTED TO MISSING" ;

PROC MEANS DATA=NHIS05.MERGED_WEIGHTED N SUM MIN MAX VAR SKEW KURT;
      VAR WTFA WTFIN_MARGINAL WTFIN_SATURATED WTFIN_CONT_AGE
      ADJFACT_MARGINAL ADJFACT_SATURATED ADJFACT_CONT_AGE;
RUN;
```

As can be seen in **Figure 3**, the original annual weight, WTFA, sums to 291,143,602 persons, and all adjusted weights do the same. WTFA has a maximum value of 19,434, and all three models have higher maximums (46,481 to 51,772), or about two-and-a-half times as much, reflecting the 50% linkage ineligibility. Variance terms for adjusted weights are larger than the original weight, also reflecting this fact. The kurtosis measures the peakedness and heavy-tailedness of a distribution; these results show that the adjusted weights are more peaked and heavier-tailed than the original weight. All of the adjusted weights are relatively close to the same value, but in the analytic phase some use of various influence statistics would be advisable. Additional code implements other checks:

```
PROC CORR DATA=NHIS05.MERGED_WEIGHTED NOSIMPLE ;
      VAR ADJFACT_MARGINAL ADJFACT_SATURATED ADJFACT_CONT_AGE ;
      WHERE WTFIN_MARGINAL>0 & WTFIN_SATURATED>0 & WTFIN_CONT_AGE>0;
RUN ;
```

```
PROC CORR DATA=NHIS05.MERGED_WEIGHTED ;
      VAR ADJFACT_MARGINAL ADJFACT_SATURATED ADJFACT_CONT_AGE ;
      WHERE WTFIN_MARGINAL>0 & WTFIN_SATURATED>0 & WTFIN_CONT_AGE>0;
RUN ;
```

We hope to see that the Pearson correlation coefficients between all of the weights are high (**Figure 4**); however, the Pearson correlation measures *linear* association, and can be misleading if nonlinearities are present, which is why we recommend examining plots as well. Likewise, we hope to see that the different modeled adjustment factors (e.g., marginal, saturated, continuous age) are highly correlated with each other. Below, we examine the plots of the weights visually to see if we can detect outliers, non linearities, heterogeneous variance, and other unusual patterns (we only display some of these plots below).

```
PROC PLOT DATA=NHIS05.MERGED_WEIGHTED ;
      PLOT WTFA*(WTFIN_MARGINAL WTFIN_SATURATED WTFIN_CONT_AGE)
      WTFIN_MARGINAL*(WTFIN_SATURATED WTFIN_CONT_AGE)
      WTFIN_SATURATED*WTFIN_CONT_AGE ADJFACT_MARGINAL*(ADJFACT_SATURATED
      ADJFACT_CONT_AGE);
      WHERE WTFIN_MARGINAL>0 & WTFIN_SATURATED>0 & WTFIN_CONT_AGE>0;
RUN ;
```

Figure 5. Plot of original public-use sample weight, WTFA, against the adjustment value, WTFIN_MARGINAL (Example 1). NHIS 2005, linkage-eligible respondents (SAS PROC PLOT).

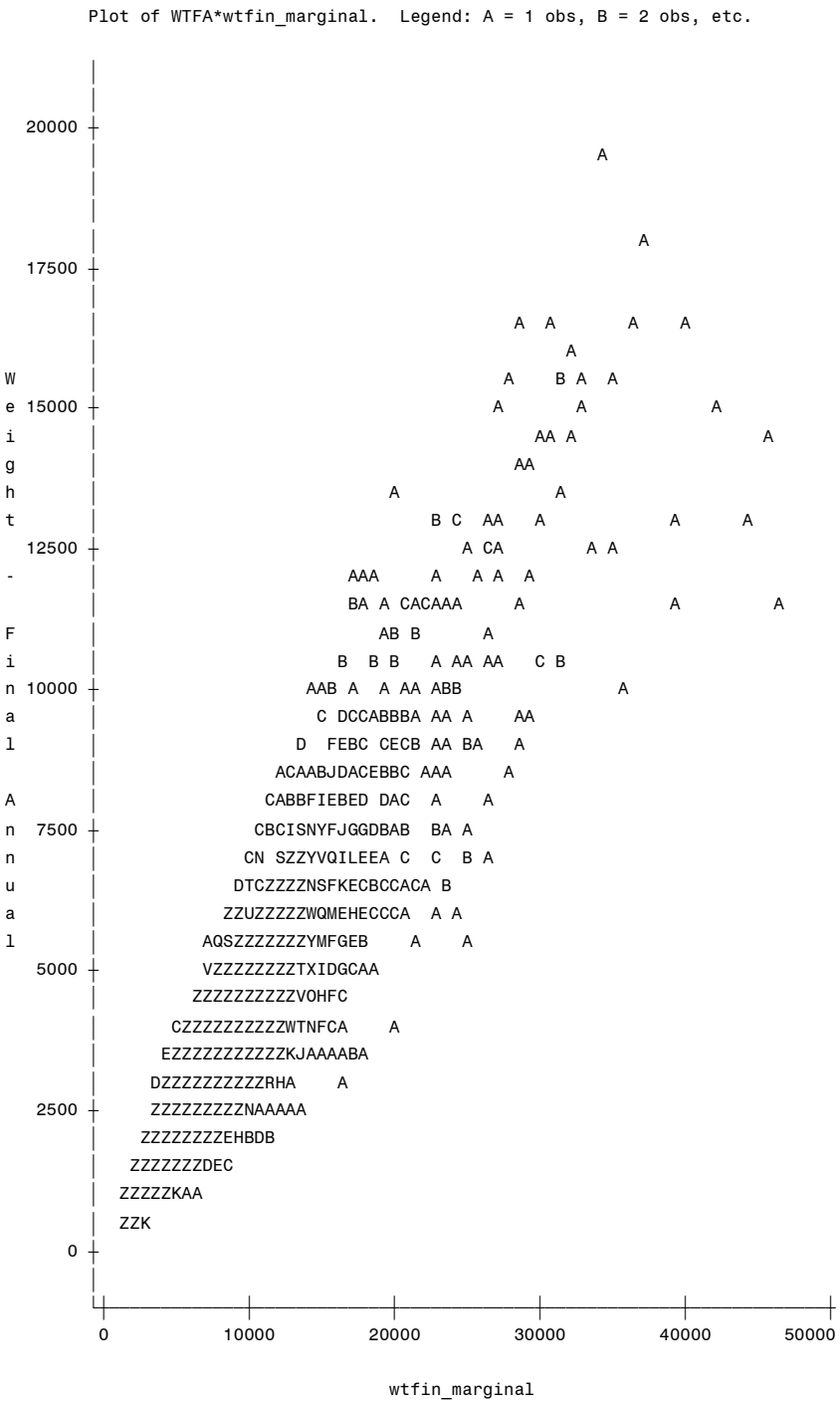
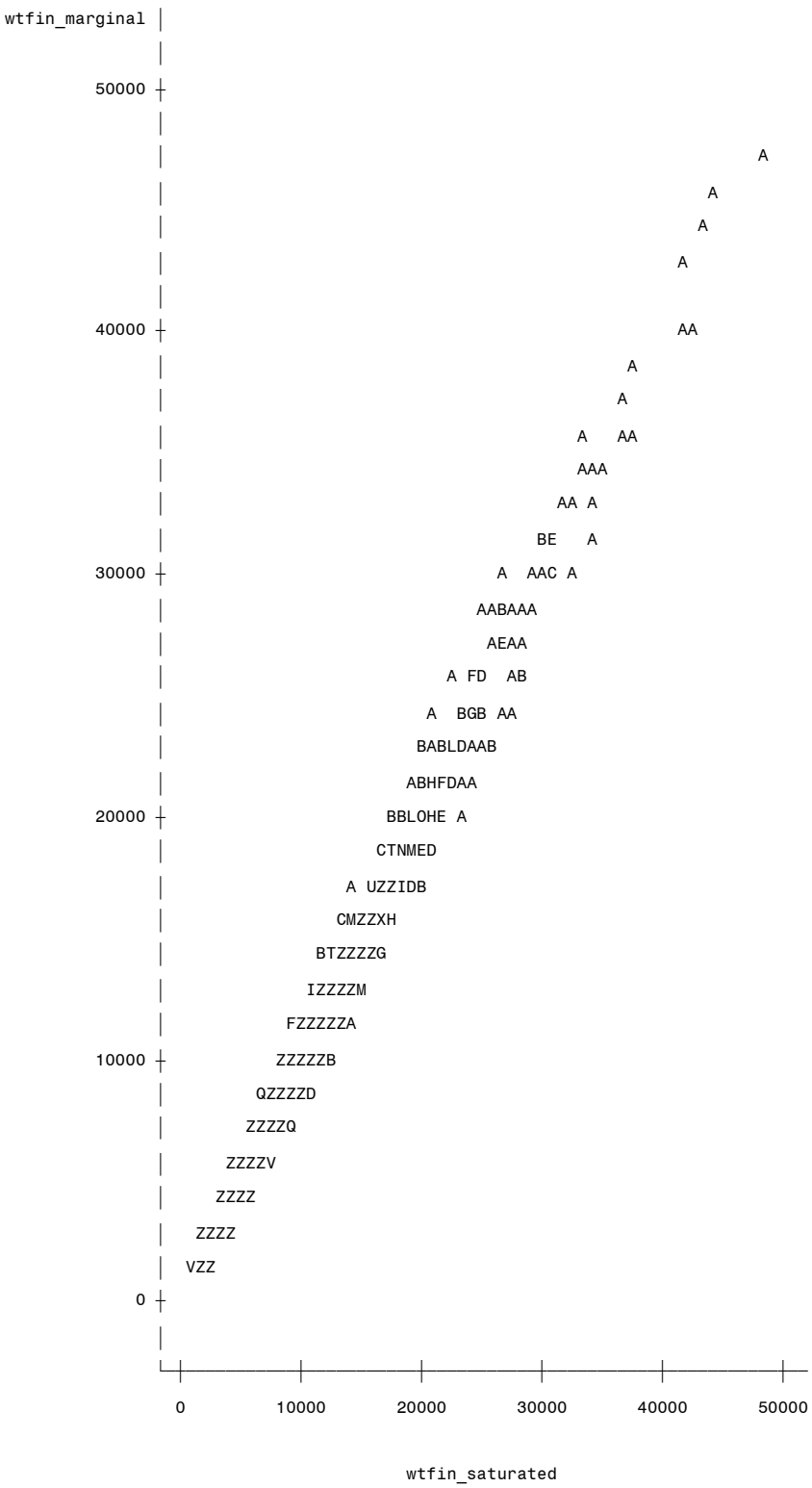


Figure 6. Plot of WTFIN_MARGINAL (Example 1) against WTFIN_SATURATED (Example 3), (SAS PROC PLOT).

```
DELETING WTFIN_MARGINAL=WTFIN_SATURATED=WTFIN_CONT_AGE=0
  Plot of wtfin_marginal*wtfin_saturated.  Legend: A = 1 obs, B = 2 obs, etc.
```




```
NOTE: 43052 obs hidden.
```

As can be seen (**Figure 5**), many values cluster in the range of zero to about 15,000, with individual values reaching about 50,000 (thus these cases may have greater influence in subsequent analytic inference). In creating these plots, the SAS convention is that overlapping values increment alphabetically (A=1, B=2, etc.). For these plots, we see that records with high weights are generally singletons and there are many more records with lower values in the lower left-hand corner. Finally, we display the plot of the marginal model weights against the saturated model weights (**Figure 6**).

The plot shows that the two very different models generate quite similar weights. This suggests that choice of weight adjustment model may not substantially affect subsequent analytic inferences.

To check for cell sizes, we need merely construct unweighted tables that reflect the properties of our reweighting model. For the marginal model, we just need to check the margins; for the model with all interaction terms (the "saturated" model), we need to check the three-way interaction. The following code implements these checks:

```
TITLE "FREQUENCY TABULATION TO CHECK CELL SIZES CELL SIZES FOR MODELS" ;
PROC FREQ DATA=NHIS05.MERGED_WEIGHTED ;
     WHERE WTFIN_MARGINAL>0 & WTFIN_SATURATED>0 & WTFIN_CONT_AGE>0;
     TABLES AGE_CAT RACEETH SEX REGION EDUC /NOPERCENT LIST ;
     TABLES AGE_CAT*RACEETH*SEX /NOPERCENT LIST ;
RUN ;
```

Examining the four margin (or univariate) tables, we found that no margin, (or category) was close to the "rule of thumb" cutoff of 30 cases (not shown). Examining the combined age/race-ethnicity/sex table, we found one minimum cell size of 51, which was still well above 30 (not shown).

Finally, we tabulated the variable FAIRPOORHEALTH, which takes on the value 1 if the respondent describes their health as "poor" or "fair", and 0 if the respondent describes their health as "good", "very good", or "excellent", and missing if missing a response (**Figure 7**). These results indicate effects of linkage ineligibility. Using the original NHIS 2005 sample and the original sample weight, 9.30% reported fair or

poor health. Even when adjusting weights in very different ways, the weighted percentages are elevated for the linkage-eligible subset: 10.43%, 10.38%, and 10.56% using the weights from the MARGINAL, SATURATED, and CONTINUOUS AGE models described in Examples 1, 3, and 4, respectively.

Figure 7. Analysis of variable “fairpoorhealth” using four sets of weights, (SAS PROC FREQ)

The FREQ Procedure

fairpoorhealth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	928009	0.32	928009	0.32
Good, Very Good, Excellent health	2.6315E8	90.38	2.6408E8	90.70
Fair/Poor health	27066765	9.30	2.9114E8	100.00

Using WTFIN_MARGINAL

The FREQ Procedure

fairpoorhealth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	295771.7	0.10	295771.7	0.10
Good, Very Good, Excellent health	2.6049E8	89.47	2.6079E8	89.57
Fair/Poor health	30357155	10.43	2.9114E8	100.00

Using WTFIN_SATURATED

The FREQ Procedure

fairpoorhealth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	296736.5	0.10	296736.5	0.10
Good, Very Good, Excellent health	2.6063E8	89.52	2.6093E8	89.62
Fair/Poor health	30216972	10.38	2.9114E8	100.00

Using WTFIN_CONT_AGE

The FREQ Procedure

fairpoorhealth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	328854.6	0.11	328854.6	0.11
Good, Very Good, Excellent health	2.6008E8	89.33	2.6041E8	89.44
Fair/Poor health	30734738	10.56	2.9114E8	100.00

References

1. National Center for Health Statistics. NCHS Data Linkage Activities. Available from http://www.cdc.gov/nchs/data_access/data_linkage_activities.htm. Accessed August 16 2012.
2. Miller D, Gindi R, Parker JD. Trends in Record Linkage Refusal Rates: Characteristics of National Health Interview Survey Participants Who Refuse Record Linkage. Paper presented at the 2011 meeting of the Joint Statistical Meeting July 30-August 4 2011 Miami Beach FL. 2011.
3. Judson, DH, Parker, JD. On dealing with "incompletely linked" data in linked survey/administrative databases: An empirical comparison of alternative methods. Paper presented at the 2012 meeting of the Joint Statistical Meetings, July 28-August 2 2012 San Diego, CA. 2012.
4. Lohr SL. "Sampling: Design and Analysis." Brooks/Cole Publishing: Pacific Grove CA. 1999.
5. Research Triangle Institute. SUDAAN Language Manual Release 10.0. Research Triangle Institute: Research Triangle Park NC. 2008.
6. Witt MB. Overview of software that will produce sample weight adjustments. Proceedings of the Section on Survey Research Methods of the American Statistical Association. American Statistical Association: Alexandria VA 3009-3023. 2009.
7. Little RJA, Rubin DB. "Statistical Analysis with Missing Data" Second Edition. John Wiley & Sons Inc.: Hoboken NJ. 2002.