Using the Sampling Design in Logistic Regression Analysis of NCHS Survey Data -- Some Applications

Van L. Parsons
National Center For Health Statistics
6525 Belcrest Rd, Room 915, Hyattsville MD, 20782

Key Words: SUDAAN, computer software

I. Introduction

The health sample surveys conducted by the National Center for Health Statistics (NCHS) have complicated design structures. Logistic regression analyses are frequently performed on these surveys, but the survey design structures are often ignored or modified, e.g., assuming independence of observations or ignoring weights. Conventional computer software, e.g., SAS PROC LOGISTIC, which cannot take into account all features of the complex design may be used. Recent software advances, e.g., Research Triangle Institute's (RTI) SUDAAN, allow the analyst to take into account design structures for a data analysis, but availability, ease of use, and cost factors often persuade analysts to continue to use conventional software. This paper compares some of the common analytical approaches to NCHS survey data.

II. Survey Design Structures

There seems to be no universal consensus as to the "correct" method for complex-survey data analysis. The concept of design-based versus model-based analysis has been discussed elsewhere (see Hansen et al. (1983) and Nathan (1988)), as have the issues involved in informative versus non-informative survey designs (see Skinner(1988) page 146).

As discussed in Korn and Graubard (1990), analysts have many options in imposing a structure on survey clustering and weighting. Many analysts feel a model-based analysis can be first run and then a common deflation factor applied to a model-based test to compensate for the survey design. Even when using the design-based structure, many analysts alter the original clustering structure and/or modify the weights. As an extreme position, the weights are ignored (i.e., set to unity), and an unweighted but clustered analysis is used. Table 1. provides some possibilities for an analytical design structure.

The with-replacement (WR) design with no clustering is usually considered to be the simplest design-based structure. Here, (w_i, x_i), $i = 1,2,\dots,n$ is a sample from a finite population and $\Pr(\text{ sample j selects element } (w_i, x_i)) \propto 1/w_i$.

Now, a model-based sample differs from a WR-design in that the data are independent, and it is usually assumed that x_i is a random variable with $\mathrm{Var}(x_i) \propto 1/w_i$.

Table 1 SURVEY DESIGN STRUCTURES

SAMPLING STRUCTURE	WEIGHTS YES	WEIGHTS NO
DESIGN-BASED		
FULL-COMPLEX DESIGN MULTI-STAGE CLUSTERS		
STRATIFIED ONE-STAGE DESIGN FIRST-STAGE CLUSTERS		
WITH REPLACEMENT DESIGN NO CLUSTERS		SRS design
MODEL-BASED		

Korn and Graubard (1990) have provided some examples and discussion of imposing various design structures on the complex survey data. They point out that even though a weighted design-based analysis is totally justifiable via sampling theory, clustering and weighting can lead to inefficiencies in the analysis. See also Kish (1992).

Many of the large scale population surveys funded by NCHS implement multi-level, disproportionate sampling strategies along with non-response and ratio adjustments. The resulting final survey weights are usually quite variable. A large sampling weight for an individual may make that individual's responses highly influential in an analysis. To avoid this, one strategy is to truncate the sampling weights, and using unweighted data is an extreme case of this strategy. After weight truncation, the population parameter estimators are biased in the design-based framework, although it is usually assumed that the mean-squared-error is reduced.

In the analytical strategies considered, the data clustering is only used in defining the functional form of the variance/covariance of an estimator. In general, if within a stratum the data clustering is ignored, the resulting variance estimator will on the average be smaller, since the data are treated as being more efficiently collected. If however, stratification is ignored and the stratification is indeed effective, the ignoring of stratification will usually increase the resulting variance. Typical variance estimators considered are based upon sums and sums of squares of cluster totals at different levels. If a design-based restructuring results in more "degrees of freedom" (df) than the original structure, then the corresponding variance estimator usually has the advantage of being more stable than the original, while if fewer "degrees of freedom" result, the variance estimator may be less stable. The modified structures will, however, often result in design-biased variance estimators. Our focus will now be directed toward logistic regression analyses of survey data under some of the various design structures discussed above.

II. Logistic Regression Analyses

To justify the modeling of a logistic regression in a finite population, a pseudo-MLE modeling is often taken, (see Skinner et al (1989) page 80 and Binder (1983)). Binder provides a detailed technical discussion, but briefly stated, one assumes that a p-vector, β, of the finite population parameters of interest is the MLE estimator for an infinite population where the finite population values are treated as the data. Furthermore, assuming that the finite population MLE solution for β satisfies the form:

$$\sum_{k=1}^{N} u(Y_k,X_k;\beta) = 0$$

where u is a p-dimensional vector valued function, (Y_k,X_k) are the N population data values, and β is the parameter to be estimated, then using the sample data, (Y_k,X_k), k = 1,2,...n, with associated weights w_k, the estimator of the population parameter is the solution to

$$\sum_{k=1}^{n} w_k\, u(Y_k,X_k;\beta) = 0 \qquad (1)$$

This is the same form as used in a weighted model-based MLE, e.g., SAS PROC LOGISTIC, thus the estimator of β, $\hat{\beta}$, is the same for both the model- and design-based structures. For the model-based approach, the variance for $\hat{\beta}$ is the inverse of the information matrix, INF^{-1}. For the design-based structure, the variance is of the form:

$$Var(\hat{\beta}) = INF^{-1}\,\Sigma_u\,INF^{-1}\,, \qquad (2)$$

where Σ_u is the variance/covariance matrix for the p-vector of expression (1). This variance linearization process for the estimator $\hat{\beta}$ is presented in greater detail in Binder (1983).

In terms of computer software, the model-based analyses are performed by conventional software, e.g., SAS, SPSS, while the design-based analyses use software specific to survey design analysis, e.g., SUDAAN, PC-CARP, OSIRIS.

In this paper the design-based analyses are done using RTI's SUDAAN (1989). This software is based on Taylor-linearization methods to estimate variances, and covers a broad class of hierarchical nested designs applicable to NCHS data. This software permits an efficient use of design information; linearized variances for means, regressions, and logistic regressions can be computed using a linear combination of Yates-Grundy-Sen variance formulas (see Cochran 1977, page 261) and usual "S^2" variance formulas. The purpose of this paper is to evaluate the effect of imposing the design structures of Table 1 on a given analysis, and not software evaluation, thus only SUDAAN has been used for the design-based structures.

SUDAAN, in its PROC LOGISTIC package, uses the methodology discussed by Binder (1983). PROC LOGISTIC provides the estimated Variance/Covariance matrix, $\hat{V}(\hat{\beta})$, for the pseudo-MLE estimated parameters, $\hat{\beta}$. SUDAAN can process both full and non-full rank design matrices and also performs some basic hypothesis tests, but its output is quite limited in scope when compared to SAS. Unfortunately, the SUDAAN procedure takes considerable mainframe computer CPU time (see Carlson and Cohen (1991) for a related study). The SUDAAN solution to the MLE equations for the logistic model requires an iteration process and multiple re-readings of the input data set. It takes over 10 times the CPU time that the same solution in SAS requires. Users of large survey databases may not have the resources needed to run SUDAAN. Thus, conventional software may be the only practical option. (This paper is based on SUDAAN version 5.53. Version 6.00 which RTI has reported as more efficient became available after the presentation of this paper. Some of the deficiencies mentioned may now be corrected.)

Most statisticians would agree that a logistic regression analysis that incorporates the complex-survey design into the sampling structure is preferable to one that does not, though there may be disagreement on the methodology to handle weighting and clustering. Our objective is to consider tests of hypotheses of the form:

Ho: $E(K\hat{\beta}) = 0$ where $K\beta$ is estimable, and K is of full row rank, with the analysis subject to the structures of Table 1.

Test statistics are based upon Wald statistics and not the -2log of the likelihood ratio, since SUDAAN is structured to produce the former. Test statistics of the following form are considered:

$$F = c * \hat{\beta}'K'(K\hat{\Sigma}K')^{-1}K\hat{\beta} \ / \ df(hypothesis)$$

where $\hat{\Sigma}$ is an estimator of the Covariance matrix associated with $\hat{\beta}$ in expression (2), c is a standardization constant, and df(hypothesis) = rank(K).

In our tables we will refer to four general types of "F" statistics :

Model-Based:

Here, the weights are scaled to $(\sum_{i=1}^{n} \frac{w_i}{n})^{-1}$ where n is the sample size. The factor c = 1.

MODEL-F : $\hat{\Sigma} = \widehat{INF}$, where if $p_i(\hat{\beta}) = (1 + \exp(-\underline{x}_i \hat{\beta}))^{-1}$ is the estimator of $Pr(Y_i = 1 \mid \underline{x}_i)$ for a logistic model

then $\hat{\mathbf{INF}} = \Sigma\, \underline{x}_i \underline{x}_i^T w_i\, p_i(\hat{\beta})\,(1 - p_i(\hat{\beta}))$.

Design Based F:

WALD-F: $\hat{\Sigma} = \hat{V}(\hat{\beta})$, with the complex design structure typically selected from Table 1 and c = 1.

For a WR imposed structure, $\hat{\Sigma}_u$ = the usual "S^2" variance/covariance matrix generated by n p-dimensional vectors $\underline{x}_i\ w_i\,(y_i - p_i(\hat{\beta}))$. This specific WALD-F will be referred to as WR-F. In general, the $\hat{\Sigma}$ matrix is generated from more complex design structures, e.g., stratum, PSU structures.

SRS-F: $\hat{\Sigma}_u$ = the weighted "S^2" variance/covariance matrix generated by n p-dimensional vectors $\underline{x}_i\,(y_i - p_i(\hat{\beta}))$, each vector having weight w_i. Equation (2) can then be used to provide an estimator for $V(\hat{\beta}$ |true unweighted SRS design). It should be noted that the $\hat{\Sigma}$'s for the WR-F and the SRS-F are identical if the weights are all unity, but in general will be different.

SAT-F: SAT-F = c * SRS-F,
c = second-order Satterthwaite correction, a function of the eigenvalues of Σ^{-1} (SRS-F) * $\hat{\Sigma}$ (WALD-F) (see Skinner (1989) page 43).

A Satterthwaite degrees of freedom will also be associated with this latter F. If the hypothesis degree of freedom = 1, then the WALD-F = SAT-F.

Under the null hypothesis, the "F-statistic" should be close to one. In practice, the F-statistic is treated as having a standard F distribution with (df_1, df_2) degrees of freedom, where df_1 = df(hypothesis), though for the SAT-F, a Satterthwaite correction is used, and df_2 = degrees of freedom associated with the estimator of Σ. For the MODEL-F, SRS-F, or WR-F the df_2 = (sample size - rank(X)), which is treated as infinite for large data sets, i.e., F is treated as having a distribution defined by a Chi-Squared variable divided by its degrees of freedom , $\chi^2(df_1)/df_1$. For a one-stage design, Stratum, PSU, df_2 is usually taken as (number of PSUs - number of strata). This is only a rule of thumb. One set of sampling conditions where the rule is reasonable is : same size PSU's, PSUs sampled with probability proportional to size, and each stratum has the same population PSU variability. In many designs the first-stage random units sampled are structurally different in different strata, and this rule cannot be applied. SUDAAN uses this rule of thumb in its probability computations. Thus, for many designs the SUDAAN p-values are not applicable to the design.

III Examples from the National Health Interview Survey.

The points discussed in I and II above will be illustrated using the 1987 National Health Interview Survey Cancer Epidemiological Supplement (NHIS-C)

The NHIS-C has about 22,000 sampled persons. The following hierarchical sampling structure seems to capture the essentials of the actual sampling mechanism.

Stratum : 52 self-representing(SR), 73 non-self-representing(NSR)

Primary Sampling Unit(PSU) : select 2 per NSR stratum using Durbin's method

Substratum: partition of PSU into at most 3 substrata for subsequent disproportionate sampling rates

Second Stage Unit (SSU): cluster of housing units within substratum (about 8 target households)

Sample household: a sample of households (usually all) is designated for interview

Sample person: one sample adult within family per household is selected for interview.

The systematic sampling mechanisms used to define sample at the second and higher sampling levels do not admit usual variance estimator formulas. The following assumptions are made to allow "classical" variance estimation formulas to be used:

1. Within substrata the SSUs result from a simple random sample with replacement and
2. All stages of sampling within a SSU will produce an unbiased estimator of that SSU's total.

Given the above framework, the following design-based structure is considered the "best" practical choice of sample design parameters with the SUDAAN software:

Full-Complex Design:

NSR strata: sample two PSUs using available Durbin individual and joint probabilities of selection, sample SSUs with replacement within Substrata of PSUs.

SR strata: sample SSUs with replacement within Substrata.

Unfortunately, confidentiality constraints, often require that some design information be omitted from the NCHS public-use data tapes. Also, some complex-design software may not allow the design detail of

SUDAAN. In such circumstances analysts are often required to impose the less efficient one-stage design:

First-Stage Design:

NSR strata: two sample PSUs are sampled with replacement within a stratum
SR strata: SSU's are randomly partitioned into two pseudo-PSU's and treated as sampled with replacement.

This latter design would usually have 125 = 73+52 degrees of freedom associated with its variance estimator, while the former design would have at least this many. As a cited rule of thumb, degrees of freedom is not applicable here since PSU's and SSU's are not comparable units. A WR-design where weights are used, but clustering ignored, will be considered for comparative purposes.

First, a logistic regression for cancer status is considered. Our study is focused upon design structures and not the relations of health and socio-economic-demographic variables, for such a study a much more detailed analysis would be required.

Response: Y = 1 (0) if the sampled individual has ever (never) been diagnosed as having cancer.

Predictors:

Class variables and levels

intercept	
sex(2):	male, female
race(2):	black, nonblack
region(4):	NE,MW,S,W
family cancer(2):	yes,no
smoking status(3):	never,former,current
industry(5):	5 classes

continuous variables: age, log(income)

The variables race, region, industry, log(income) should be highly correlated with the NHIS geographic stratification variables.

This is a non-full rank model; there are 21 parameters, but the rank of the X matrix is 15. The comparisons of the different designs structures are in Table 2. In this table the F-tests are testing the hypothesis H_0: specified level parameters = 0.

Impact of Analysis Structure on Inference:

Analyses using .05 level tests are considered. Since the sample size is quite large, the "F-statistic" values for the WR, SRS and the model-based designs are compared to a chi-squared distribution divided by its degrees of freedom. For the Full complex or First-Stage designs, the SAT-F and WALD-F it is really not clear whether the rule of thumb denominator degrees of freedom 125 is reasonable. We suggest considering:

$$\text{minimum } \{df_2: P(F(df_1,df_2) > \text{observed }) \leq .05.\}$$

for a given test. If this value is "small" the analyst may feel confident the F-test is significant.

An examination of Table 2 shows that there is a strong positive correlation of the various F-statistics. For all structural designs and weighting one would have

Not significant: Industry
Significant: Age, Race, Sex, Region, Family Cancer, Smoking Status

Inference on Log(income) is mixed. For an un-weighted model design or unweighted WR-F = SRS-F the observed F would be judged significant since $P(F_{1,\infty} > 3.85) = .05$.

Some comments on the NHIS-C analyses

1. Relative comparisons among the different structures show a lack of uniformity in magnitude. No one structure gives larger or smaller F statistics.

2. As one might expect, for a fixed weight structure, the Model-F and SRS-F's seem reasonably comparable.

3. For the Full and First-Stage design options, the SAT-F is considered to be more stable than the WALD-F and is the preferred design-based test statistic. In most of these cases we see that the SAT-F is of smaller magnitude than the WALD-F. If SAT-F > SRS-F then estimated design effects (DEFT = ratio of complex to SRS-sampling standard errors) may be less than one. A small estimated deft may be the result of instability of the complex design variance estimator, but it may also indicate insufficient data. There is also the possibility that the survey stratification was extremely effective, but this seems rare in population surveys. The weighted estimate for the race effect had DEFT < 1. An examination of the sampling weights associated with the black-male sample showed a large amount of variability with the bulk of cancer cases concentrated on the smaller weights. This distribution may contribute to the small DEFT. For the unweighted design the DEFT is greater than unity.

4. The First-Stage design tend to produce larger variance estimators than the Full-Complex design. Most of the F statistics for the former are smaller.

5. The SAT-F value tends to be smaller than the MODEL-F or SRS-F for a fixed design.

6. The WR-F and SRS-F statistics are biased estimators for $V(\hat{\beta} \mid$ true complex design) but should be more stable than the WALD-F statistics based upon clustered data. Except for the test of Regression the weighted WR-F was close to SAT-F(Full-Complex).

In the next NHIS-C example we consider two variables which should be independent of the design structure but also have similar distributional properties: month of birth January and month-of-birth March. The predictors used are age, sex, race, region, poverty and education as in the first example. For the response month-of-birth January the over-all test for regression was not significant, but for month of birth March, the overall test of regression could be argued to be significant between the .01 and .05 levels using the various structures, (see Table 3). Regional and Poverty parameters may also be judged significant. While such observations can be attributed to chance, it must be kept in mind that the data are also subject to non-sampling errors which we have little knowledge or control. Extra care must be taken in interpreting marginally significant results.

From these NHIS examples, it seems that one major difficulty in testing the hypotheses is determining a denominator degrees of freedom so that the F-statistics can be assessed from an F-table. The various F-statistic are highly correlated, if one F is of "large" magnitude then most likely they all are. Inference made on "large" magnitude F-statistics is probably sound. Making decisions on a perceived moderately sized p-value, say .01 to .05, with survey data seems risky since underlying assumptions of "large sample" theory may not actually hold.

Recommendations

A complete analysis of survey data involves some justification of the imposed probabilistic structure on the data. A design based analysis using the weights and clustering is the only one that can easily be defended since it is a consequence of the actual sampling mechanism. The SAT-F is the preferred statistic for the design-based approach. Highly modified design-based approaches would have to be justified with respect to biases resulting from restructured clustering and weighting. The model-based analysis assumes a controlled experiment. The justification of this methodology seems more complicated. Some design-based analysis recommendations are now presented.

1. Perform some diagnostics on the design weights and the clustering. It might be reasonable to truncate the largest sampling weights somewhat. The resulting additional design-based bias may be traded-off against increased stability of the variance/covariance.

2. Use a conceptional design that is efficient as possible.

3. Use the Satterthwaite adjusted F statistic.

4. In an analysis one should assess the fit of the logistic regression model. As discussed in Hosmer et al (1991) and Pregibon (1981), measures of residual, leverage and influence should be considered. Many of the Pregibon diagnostics are programmed in SAS PROC LOGISTIC for the model-based approach. Since $\hat{\beta}$ is the same as in SUDAAN, these diagnostics may be informative, even for the design-based approach.

5. Examine the design effects for small values which may indicate insufficient data.

6. P-values reported using large sample theory may be suspect.

7. Do the analysis using several of the mathematical structures of Table 1. Conflicting results should require further study.

References:

Binder, D.A. (1983) On Variances of Asymptotically Normal Estimators from Complex Surveys, *International Statistical Review*, 51 279-292.

Carlson, B.L. and Cohen, S.B. (1991) Evaluation of the Efficiency of Using Personal Computers for Regression Analysis on Complex Survey Data, American Statistical Association 1991 *Proceedings of the Section on Survey Research Methods*, to appear.

Cochran, W.G. (1977) *Sampling Techniques*, 3d ed., John Wiley and Sons, New York.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983) An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys, *Journal of the American Statistical Association*, Vol. 78, pages 776-793.

Hosmer, D.W., Taber, S., Lemeshow, S., (1991) The Importance of Assessing the Fit of Logistic Regression Models: A Case Study, *American Journal of Public Health*, Vol. 81, pages 1630-1635.

Kish,L, (1992) Weighting for Unequal P_i, *Journal of Official Statistics*, Vol. 8, pages 183-200.

Korn, E.L. and Graubard, B.I. (1991) Epidemiologic Studies Utilizing Surveys: Accounting for the Sampling Design, *American Journal of Public Health*, Vol. 81, pages 1166-1173.

Massey, J.T, Moore, T.F., Tadros, W., and Parsons, V.L. (1989) Design and estimation for the National Health Interview Survey, 1985-1994. National Center for Health Statistics (1973) Plan and Operation of the Health and Nutritional Examination Survey, United States, 1971-73, *Vital and Health Statistics*, Series 1-Nos 10a and 10b, DHEW Pub. No. (HSM) 73-1310, U.S. Government Printing Office, Washington.

Nathan, Gad, (1988) Inference Based on Data from Complex Sample Designs, *Handbook of Statistics*, Vol 6, Krishnaiah and Rao, eds,Elsevier, pages 247-266.

Pregibon, D., (1981) Logistic Regression Diagnostics, *Annals of Statistics*, vol. 9, pages 704-724.

SAS (1985) *SAS User's Guide: Statistics*, Version 5 Ed. Carey,NC:SAS Institute.

Skinner, C.J., Holt, D., and Smith, T.M. (1989) *Analysis of Complex Surveys*, John Wiley and Sons, New York.

SUDAAN (1989) *Software for Survey Data Analysis*, Research Triangle Institute, North Carolina.

Table 2 NHIS DATA

Weighted (W) and Unweighted (U)
Full-Complex Design (D1) and
First-Stage Design (D2)
With-Replacement Design (WR)

RESPONSE = CANCER STATUS

TEST	DF	SAT DF	WALD F	SRS F	SAT F	MODEL F
REGRESS						
W-D1	14	13.0	60	72	54	58
W-D2		11.4	63		42	
W-WR			60			
U-D1		13.5	73	70	68	58
U-D2		12.3	80		58	
AGE						
W-D1	1	*	*	595	501	556
W-D2		*	*		392	
W-WR			474			
U-D1		*	*	571	592	542
U-D2		*	*		510	
SEX						
W-D1	1	*	*	45.2	33.1	44.4
W-D2		*	*		29.7	
W-WR			34.3			
U-D1		*	*	41.3	44.0	41.2
U-D2		*	*		36.4	
RACE						
W-D1	1	*	*	24.5	31.4	24.8
W-D2		*	*		31.9	
W-WR			30.4			
U-D1		*	*	26.1	25.0	26.3
U-D2		*	*		22.1	
REGION						
W-D1	3	2.9	11.9	16.1	11.5	16.3
W-D2		2.7	10.8		8.6	
W-WR			12.8			
U-D1		3.0	12.4	12.3	12.5	12.5
U-D2		2.8	12.3		11.1	
FAMILY CANCER						
W-D1	1	*	*	27.8	22.1	28.8
W-D2		*	*		19.0	
W-WR			21.9			
U-D1		*	*	37.6	40.7	38.8
U-D2		*	*		34.8	
SMOKE STATUS						
W-D1	2	2.0	19.6	21.3	17.5	21.4
W-D2		2.0	19.8		17.0	
W-WR			16.8			
U-D1		2.0	22.7	20.5	20.2	20.6
U-D2		2.0	24.5		20.5	

Table 2 NHIS DATA (continued)

TEST	DF	SAT DF	WALD F	SRS F	SAT F	MODEL F
INDUSTRY						
W-D1	4	3.9	1.8	2.2	1.7	2.3
W-D2		3.8	1.8		1.5	
W-WR			1.9			
U-D1		4.0	0.85	0.87	0.86	0.91
U-D2		3.9	0.91		0.83	
LOG(INCOME)						
W-D1	1	*	*	2.0	1.7	2.1
W-D2		*	*		1.4	
W-WR			1.7			
U-D1		*	*	4.0	3.5	4.3
U-D2		*	*		3.2	

* WALD-F = SAT-F SAT DF = 1

Table 3 NHIS DATA

RESPONSE = BORN IN MONTH OF MARCH

Full-Complex Design (D1)

	DF	SAT DF	WALD F	SRS F	SAT F	MODEL F
REGRESS						
W-D1	9	8.7	2.6	3.2	2.5	3.1
U-D1		8.8	2.6	2.5	2.4	2.5
AGE						
W-D1	1	*	*	1.4	1.1	1.4
U-D1		*	*	1.3	1.2	1.2
SEX						
W-D1	1	*	*	0.18	0.12	0.18
U-D1		*	*	0.37	0.32	0.37
RACE						
W-D1	1	*	*	0.73	0.61	0.74
U-D1		*	*	0.12	0.12	0.12
REGION						
W-D1	3	3.0	4.7	6.0	4.7	6.0
U-D1		3.0	4.7	4.5	4.5	4.5
POVERTY						
W-D1	1	*	*	4.0	4.1	4.1
U-D1		*	*	4.9	5.0	4.9
EDUC						
W-D1	2	2.0	2.1	2.5	2.2	2.5
U-D1		2.0	2.0	1.8	2.0	1.7

* WALD-F = SAT-F SAT DF = 1