Regression Inference on the National Health Interview Survey- An Empirical Study

Van L. Parsons
National Center For Health Statistics, Hyattsville MD, 20782

Key Words: complex-survey, SUDAAN

I. Introduction

The National Health Interview Survey (NHIS) is an annual survey which monitors the health of the Nation. In analysing data from this survey, regression analyses are often performed to provide insights about the population. In the past, the lack of commercially available computer software for complex-survey design data had either limited the scope for such analyses or forced the analyst to use regression software intended for a controlled experiment, e.g., SAS PROC GLM or PROC REG.

Recently, Research Triangle Institute (RTI) has developed commercially available software for the analysis of data from complex surveys: SUDAAN (1989). This software uses Taylor-linearization methods to estimate variances, and extends the previous RTI packages (SESUDAAN, SURREGR, RATIOEST) to include a broader class of hierarchical nested designs. The old RTI software was restricted to designs supporting simple random sample (SRS) within each level. In particular, SURREGR, the program for regression analysis, required an assumption of only one level of sampling; the variance estimator was based upon the functional form of the usual estimate of SRS variance. The new software allows for more efficient use of design information; variances for the NHIS can now be computed using a linear combination of Yates-Grundy-Sen variance formulas (see Cochran 1977, page 261) and SRS variance formulas.

While regression software for complex survey designs is available, it may be more cumbersome and/or costly to run than software based upon controlled experiment data. Since the functional form of regression parameter estimators are identical for a weighted least squares (WLS) regression and the complex-survey design regression, many analysts feel a WLS can be first run and then a common deflation factor applied to all "F-tests" to compensate for the survey design. Also, some analysts feel that the weights can be ignored (i.e., set to unity), and an ordinary least squares (OLS) regression can be used to find relations among response and predictor variables. See Korn and Graubard (1991) for related discussion.

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 regression analysis has been discussed elsewhere (see Nathan (1988)), as have the issues involved in informative versus non-informative survey designs (see Skinner(1988) page 146).

This paper presents NHIS data analyses using some commonly assumed data structures. The work presented here is not intended to be comprehensive in scope. But the data presented here was in response to the many inqueries NCHS receives on how to analyze the NHIS over time. The data sets discussed tend to have large samples of individuals, but the data tend to be highly clustered, much more so than in a cross-sectional analysis of an annual survey.

In the following, regression analyses are performed using the NHIS design structures and comparisons are made to models with designs applicable to non-complex-survey software, in particular, OLS and WLS. Typical NHIS realizations are simulated over a 10 year period, and regression analyses using different conceptual models are considered.

II. Conceptual 1985-1995 NHIS Sample Design

A detailed description of the NHIS design is given in Massey et al.(1989), but for practical purposes, the design can be conceptualized in a classical hierarchical framework. A brief outline is now presented.

The strata/sampling units of the multistage NHIS Supplement have the following hierarchy:

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 parts for differential sampling rates (for subdomains of interest)

Survey-segment: cluster of housing units to cover 10 years of sample; the entire unit is selected systematically within substratum

Annual cluster: a random mechanism defines 10 subclusters (of about 8 households each) with the survey-segment for each annual survey

Sample household: a sample of households (usually all) is designated for interview within each 1 year cluster

Sample person: for the Core NHIS, all persons within a household are interviewed; for a typical Supplement, one sample adult 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. It is assumed that the sample survey-segments are the result of a simple random sample with replacement (SRS) and

2. It is assumed that all stages of sampling within a survey-segment will produce an unbiased estimator of annual cluster total for any year or aggregate of years.

Given the above framework, a simplified design structure is to consider the NHIS as having 2 sampling levels: first, select PSUs, and then select survey-segments. Each sample survey-segment has an unbiased estimator of total for individual years or aggregate years, and from these components the usual Horvitz-Thompson estimator of population total can be computed. This simplified design will probably lead to slightly conservative estimators of variance. Also, as mentioned in the introduction, the concern is with combined years of NHIS data. To work within the constraints of the NCHS computer system (CPU time, memory) the study was restricted to NHIS Supplement designs.

III. Regression Analyses

To justify the modeling of a regression in a finite population, it is often postulated that the vector of population values, $\underline{Y}$, can be expressed as the expectation of a super-population: $E\underline{Y} = X\beta$ where β is a p-vector of unknown parameters, and X is a p-column population design matrix. The population least squares estimator of β is

$b = (X'X)^{-}X'\underline{Y}$, and the sample survey estimator is

$b = (X'WX)^{-}X'W\underline{Y}$, with W a diagonal matrix of weights, where the X and $\underline{Y}$ are the sample realizations. (If rank X < p, then we have a non-full rank model. In this case, generalized inverses are used, and b is referred to as a solution.)

The estimator b under the finite population sampling approach is identical to the WLS estimator, but distributional properties of b depend upon the probabilistic structure imposed by the sample design. Two recent references on complex sample survey regression are Skinner, Holt and Smith (1989) and Kott (1991).

Typically, survey data is analyzed using one of the analysis structures of Table 1. Analyses may be performed weighted or unweighted; in the latter case, the matrix W is replaced by the identity matrix I. The variance-covariance matrix for b, V(b), depends upon the imposed sampling structure. For model-based analysis, homogeneous pure error assumptions result in WLS or OLS estimates of variance typically produced in standard computer software like SAS PROC GLM or PROC REG. For the complex-survey and with replacement designs, the estimator b is linearized

$$b_L = (X'WX)^{-} \Sigma w_i \underline{x}_i (y_i - \underline{x}_i'b),$$

where $\underline{x}_i$ = column of predictors, and then this new variable is substituted into the variance formula for an estimator of total. Complex-survey designs treating the weights as unity are discussed in Korn and Graubard (1991), but are not discussed here.

A with-replacement (WR) design assumes a heteroscedatic error structure of the y_i's under sampling with replacement. If W = I, then the variance estimator of b corresponds to the SRS linearization estimator described in equation (3.21) of Skinner et. al. (1989). The SAS procedure REG (see page 660 of SAS (1985)) can compute such a WR-design variance, though it seems PROC GLM does not have this option. The SUDAAN version 5.50 software procedure REGRESS computes an estimate of V(b) for a broad class of hierarchical nested designs. For the NHIS complex design, the variance formula for total is described in Massey et al. (1989), equation(7), page 32; this is generated by the SUDAAN software.

Proc REGRESS in SUDAAN provides the estimated V(b) matrix for both full and non-full rank design matrices and also performs some basic hypothesis tests, but its output is quite limited in scope. It does not have all the special options of SAS procedures GLM and REG. As such, the analyst must often do additional processing outside the SUDAAN environment. Furthermore, the SUDAAN procedures take considerable mainframe computer CPU time (see Carlson and Cohen (1991)). Mainframe usage constraints of cost and time will restrict many analysts from intensive SUDAAN runs on large data sets. Because SUDAAN is still in the development stage, many future improvements may be anticipated.

Most statisticians would agree that a regression analysis which incorporates the complex-survey design into the sampling structure is preferable to one that does not, though there may be disaggreement on the methodology to handle weighting and clustering. In this paper the impact of imposing the simplified

sampling structures of Table 1 on data from an NHIS survey is considered. Our general objective is to consider tests of hypotheses of the form

Ho: E(Kb) = 0 were Kb is estimable, and K is of full row rank.

The generic "F-test" statistic to be considered is

$$F = c * b'K'(K\Sigma K')^{-}Kb / df(hypothesis)$$

where Σ is a Covariance matrix , c is a constant, and df(hypothesis) = rank(K)

In our tables we will refer to "F" statistics, defined according to the imposed survey structure:

Design Based:

WALD-F: Σ = V(b | full complex design), c = 1

WR-F: Σ = V(b | simplifed WR design), c = 1

SAT-ADJ-F: Σ = V(b | simplifed WR design), c =second-order Satterwaith correction, a function of the eigenvalues of inv(V(b | WR design)) * V(b | complex) (see Skinner (1989) page 43)

Model-Based:

IID-F : Σ = V(b | WLS homogeneous error model), c =1

The Covariance matrix, Σ, will be estimated from the data.

SUDAAN produces WALD-F and SAT-ADJ-F for some select K. More importantly, SUDAAN can create an output file consisting of b, estimates for Var(b | full complex), Var(b | simplified WR design), (XWX), and its generalized inverse, Ginv(XWX). In this paper these output matrices were used to compute (outside the SUDAAN environment) all the "F" statistics in our tables. Note, the homogeneous pure error variance can be obtained from the V(b | simpified WR design) output matrix, thus, SAS PROCs GLM or REG were not used for any of the runs. The model-based results were confirmed by cross-checking with these SAS procedures on some smaller data sets.

IV Examples.

The data for the examples came from 3 combined years, 1987-1989, of an NHIS supplement (one adult per household) sample for persons aged 65 and older. This data set had 21,369 observations. We considered the regression:

Response: Y = Square Root(number of doctor visits in year)
Predictors:

Intercept

Class variables and levels - year(3), region(4), poverty status(2), sex(2), race(2), (sex*race)(4)

Continuous variables- age, (ratio of body weight to height), size of psu, and
combined linear + quadratic variables for family income, education and family size

This is a non-full rank model; there are 27 parameters, but the rank of the X matrix is 19. The results of the different regressions are in Table 2.

Comments on Table 2:

1. Except for levels Sex and Race, the F-tests are testing the hypothesis H_0 : specified level parameters = 0.

For Race or Sex, a test of hypothesis is made on a reduced-model where the interaction (Race*Sex) = 0 by constraint. The reduced population parameter of equation (103) of Searle (1971) was considered and then estimated using the sample data. Of course, in the presence of significant interaction, the analyst would disregard the reduced test as meaningless. These F tests correspond to the SAS Type-II F tests.

2. If the hypothesis degree of freedom = 1, then the Wald-F = SAT-ADJ-F.

3. IF the null hypothesis is true, the F statistic should be close to one. In a controlled experiment with homogeneous normal errors, we have the IID-F has an F distribution with (hypothesis, sample size - rank(X)) degrees of freedom. For this complex design it is not clear what denominator degrees of freedom (df) should be associated with the estimator of V(b | complex design). Often, the rule of thumb is df = (number PSUs - number of strata). In our survey each NSR stratum would have 1 associated degree of freedom, but it would be wrong to define one PSU per SR stratum giving 0 df's. It would also be misleading to use the number of sample survey-segments within the SR strata, about 3000. Unfortunately, SUDAAN computes denominator degrees of freedom solely on the listing in the NEST statement; our specification would result in the 52 SR strata each having 0 df, and consequently, the p-values computed in SUDAAN might be invalid.

4. In this example, all the V(b) matrices had about the same level of stability whether we used the complex, WR or IID designs. The stability ratio:

SQRT(min(d_i) / mean(d) * (r+1)/2) where $d_1, d_2, \ldots, d_r$ are the eigenvalues of $V_f(b)$, the full rank submatrix of rank r of V(b), was about .08 for each of the designs we ran. We experienced no numerical difficulty with matrix inversion needed for the tests.

Impact of Analysis Structure on Inference:

Analysis using .05 level tests were considered. Since the sample size is so large, the "F" values for the WR designs and the IID model designs were compared to a chi-squared distribution. For the SAT-ADJ-F and Wald-F it is really not clear what value to use for denominator df (den-df), so that the minimum df for the denominator required to reject the hypothesis, i.e., minimum {den-df : P(F(num-df, den-df) > observed) $\leq$ .05.} was computed.

For example, the SEX*RACE interaction measure of 4.2 would be significant for the SAT-ADJ-F if the den-df $\geq$ 28. The significance or non-significance of the following predictors would most likely be judged the same regardless of imposed design structure:

Not significant: Year, Income, Poverty
Significant: Age, (Race,Sex,Interaction), Size of PSU

Inference on the other predictors is mixed. The inferred models will be (denoting C as the common values, significant for all design structures)

SAT-ADJ-F	Y = C
WR-F (weighted)	Y = C + (Family Size)
IID-F (weighted)	same as weighted
WR-F (unweighted)	Y = C + (Wt/HT) + (Region) + (Education)
IID-F (unweighted)	same as weighted

Caveats:
(1) These results were produced without checking outliers, and the usual diagnostics which one performs in a data analysis.

(2) Under a complex design, the exact distribution of the "F" statistics are unknown.

The aged 65+ regression analysis demonstrates that the different analyses result in different inferences about the population, but questions of the actual significant level and power of the tests cannot be answered. A simulation would require a generation of samples at all levels of sampling from an appropriate universe file. Such a project would be quite involved.

V. Simulation of data for a 10 year Supplement

To better understand the impact of the clustering effect on regression analyses over time, 10 years of NHIS Supplement data were generated. To keep the problem manageable, the simulation was restricted to one of the four regions of the U.S. (West) rather than the entire U.S. A total of 8 data sets were generated using the following model: From the 1987 NHIS sample PSUs, survey-segments, and households were fixed. The households were indexed by race and sex of a randomly chosen adult. For each of 10 years, a household response of random household size (using a race distribution) and a response with distribution a function of sex and year:

E(response) = μ(sex) + .01*μ(sex) * (year-1),
Var(response) = σ^2(sex) was generated.

A parallel profile model by year and sex as generated. No race or (race*sex) factor was present. Responses consisted of 4 normal and 4 gamma variables, N1-N4 and G1-G4 respectively, where the index represents increasing correlation within the survey-segments. Note: these realizations will have less variability over time than would be observed in practice.

Simulation results: The design structure is identical to that of aged 65+ example discussed previously, but now we have about 100,000 observations, 22 SR Strata, 11 NSR Strata with 2 PSUs each, and 1600 sample survey-segments. The regression model run has predictor variables: year, race, sex, and (sex*race). The types of tests run are comparable to those discussed in the aged 65+ example. The tests for (sex*race) interaction, race(reduced model with no interaction), and year are presented for the 8 different simulated variables in Tables 3-5. The test for sex was highly significant for all analysis structures and is omitted from the tables.

Comments:

1. For all 8 runs, the sex*race interaction would be judged insignificant (at level .05) when making inference using the complex survey SAT-ADJ-F. For simulations N3 and N4, the IID-F and WR design gives significant results, both weighted and unweighted for N3, and unweighted for N4. The variables N3, N4, G3, G4 were generated to be correlated within survey-segment. The hypothesis of an independent error structure in the analysis is violated in the WR and IID models.

2. In the presence of no sex*race interaction, one often tests no race effect subject to sex*race interaction = 0. The SAT ADJ F tests were all insignificant, but N4 which had insignificant weighted IID-F and WR-F statistics had a significant race effect.

3. The test for year effect will be significant in all cases so long as we could assume 10 or more degrees of freedom associated with the estimated

Variance/covariance matrix.

VI. Conlusions: From the examples, it is apparent that the design and weighting structure imposed upon a complex survey for data analysis may affect the resulting inference. It appears that imposing the full complex design structure on the data makes rejection of a null hypothesis more difficult than the other design structures. When the data exhibits a large amount of clustering, the IID and WR methods tend to under-estimate variance, and the "F" tests may be misleading. Data analysts should apply simplified design structures to complex survey data with caution.

References:

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.

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

Kott, P.H. (1991) A Model-Based Look at Linear Regression with Survey Data, *The American Statistician*

TABLE 1
ANALYSIS STRUCTURES

	DESIGN BASED		MODEL BASED
WEIGHTS	COMPLEX SURVEY DESIGN	WITH REPLACEMENT DESIGN	HOMOGENEOUS PURE ERRORS
W	SUDAAN (UNEQWOR)	(WR)	SAS (WLS)
W=I		(SRS)	SAS (OLS)

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, *Vital Health Stat* 2(110).

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

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

Searle, S.R., (1971) *Linear Models*, John Wiley and Sons

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
F STATISTICS

	LEVEL	DF Hyp	DF SAT	COMPLEX DESIGN WALD F	WR DESIGN WALD F	COMPLEX DESIGN SAT ADJ F	MODEL I.I.D. F
W	Regression	18	15.5	10.0	12.4	7.6	12.9
U		18			11.7		12.7
W	Year	2	1.9	2.1	2.9	2.1	3.0
U		2			1.2		1.2
W	Age	1	1.0	79	108	79	114
U		1			63		67
W	Sex	1	1.0	18	26	18	28
U		1			35		37
W	Race	1	1.0	5.5	7.0	5.5	10.9
U		1			12.7		16.7
W	Sex*Race	1	1.0	4.2	4.9	4.2	8.1
U		1			10.4		14.6
W	Wt/Ht	1	1.0	1.7	2.0	1.7	2.3
U		1			6.8		8.0
W	Income	2	2.0	1.0	1.3	1.0	1.4
U		2			2.2		2.3
W	Education	2	1.8	1.9	1.3	1.2	1.4
U		2			3.4		3.7
W	Family Size	2	1.8	2.7	4.1	1.8	4.5
U		2			1.4		1.5
W	Size PSU	1	1.0	6.4	11.0	6.4	10.3
U		1			14.5		13.7
W	Region	3	2.9	1.3	2.5	1.5	2.4
U		3			2.8		2.7
W	Poverty	1	1.0	0.1	0.1	0.1	0.2
U		1			0.9		1.0

TABLE 3
TEST : NO SEX*RACE INTERACTION

F STATISTICS

VARIABLE		DF HYP	DF SAT	COMPLEX DESIGN WALD F	WR DESIGN WALD F	COMPLEX DESIGN SAT ADJ F	MODEL I.I.D. F
G1	W	1	1	0.0	0.0	0.0	0.0
	U	1			0.1		0.1
G2	W	1	1	0.3	0.5	0.3	0.5
	U	1			0.8		0.7
G3	W	1	1	1.1	1.8	1.1	1.7
	U	1			1.6		1.4
G4	W	1	1	1.2	3.2	1.2	3.0
	U	1			0.8		0.7
N1	W	1	1	0.2	0.2	0.2	0.1
	U	1			0.1		0.1
N2	W	1	1	0.0	0.0	0.0	0.0
	U	1			0.3		0.3
N3	W	1	1	3.2	5.4	3.2	5.0
	U	1			5.1		4.7
N4	W	1	1	1.0	2.5	1.0	2.3
	U	1			8.0		7.3

TABLE 4
TEST : NO RACE DIFFERENCE

F STATISTICS

VARIABLE		DF HYP	DF SAT	COMPLEX DESIGN WALD F	WR DESIGN WALD F	COMPLEX DESIGN SAT ADJ F	MODEL I.I.D. F
G1	W	1	1	0.7	0.7	0.7	0.6
	U	1			0.0		0.0
G2	W	1	1	1.6	2.8	1.6	3.0
	U	1			0.3		0.3
G3	W	1	1	0.3	0.7	0.3	0.7
	U	1			0.5		0.5
G4	W	1	1	0.4	1.2	0.4	1.2
	U	1			2.7		2.7
N1	W	1	1	2.0	2.1	2.0	2.1
	U	1			0.7		0.8
N2	W	1	1	0.8	1.1	0.8	1.1
	U	1			2.9		2.9
N3	W	1	1	1.4	2.4	1.4	2.5
	U	1			1.5		1.5
N4	W	1	1	1.8	4.1	1.8	4.2
	U	1			5.3		5.5

TABLE 5
TEST : NO CHANGE BY YEAR

F STATISTICS

VARIABLE		DF HYP	DF SAT	COMPLEX DESIGN WALD F	WR DESIGN WALD F	COMPLEX DESIGN SAT ADJ F	MODEL I.I.D. F
G1	W	9	7.8	5.0	5.5	3.5	5.4
	U	9			5.7		5.8
G2	W	9	8.5	3.0	4.2	3.2	4.2
	U	9			5.8		5.9
G3	W	9	7.4	3.5	4.8	3.2	4.9
	U	9			4.3		4.4
G4	W	9	7.1	3.6	5.1	3.3	5.1
	U	9			3.6		3.6
N1	W	9	8.5	4.2	5.2	4.0	5.2
	U	9			4.4		4.5
N2	W	9	8.4	5.3	7.9	6.1	7.9
	U	9			6.3		6.2
N3	W	9	8.0	4.1	5.2	3.5	5.1
	U	9			5.4		5.4
N4	W	9	8.3	5.1	4.6	3.9	4.6
	U	9			4.2		4.2