Evaluation of Model-based Methods in Analyzing Complex Survey Data: A Simulation Study using Multistage Complex Sampling on a Finite Population[1]

Van Parsons[2], Rong Wei[2], Jennifer Parker[2]
[2]National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Rd, Hyattsville, MD 20782

Abstract

The usage of traditional design-based methods for complex-survey data often leads to estimation difficulties or unreliable analyses when sample sizes are not sufficiently "large" at some level of multi-stage sampling. In these situations model-based estimation methods are often suggested as alternatives to compensate for data deficiencies. For this study, we focus on both design- and model-based statistical inference based on a sample of ~1000 households taken from a reduced-scale pseudo U.S. population that captures many features of geographical and household clustering within the true U.S. population. This pseudo population was developed from nine years of the National Health Interview Survey (NHIS) data. A simulation study is performed on this pseudo-universe where we imposed a complex design including multilevel sampling from strata, primary clusters, secondary clusters, households, and persons along with post-stratification weighting adjustments. Sampling properties of design-based regression estimators (Binder 1983) and multi-level model-based regression estimators are compared.

Key Words: random effects, multilevel, degrees-of-freedom

1. Introduction

Analysts frequently use population-based, multi-stage complex surveys to make population inferences or to make inferences about relationships among variables. The National Health Interview Survey (NHIS), Botman et al. (2000), is such a survey; it is rich in information about health and related variables for the U.S. population, and its data are targeted for health related analyses. The analysis of complex survey data is often classified by two types of "randomness" imposed upon the survey data. The *design-based* analysis assumes that randomness is based on all possible sample selections, while the *model-based* analysis imposes well-defined distributional structures on the sample. Pfeffermann (2011) discusses the nature of complex survey data and the motivations and limitations of design-based and model-based methods in some detail. The NHIS' operating characteristics follow those discussed in that paper.

The focus of this paper is upon suggesting and justifying analytic strategies for the analysis of the NHIS data as practiced by typical NCHS data users. While design-based strategies are generally accepted as providing reliable inferences when the analyses focus on estimates of population means over large national domains and the assumption of

[1] *The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control*

asymptotic "normal" distribution sampling behavior is taken as reasonable, many NHIS data analyses are now targeting smaller domains and the associations of multiple covariates with health variables. As analyses become more sophisticated, the necessary (asymptotic) conditions for obtaining reliable design-based estimates often become difficult to achieve. For example, due to the coarse nature of cluster sampling (especially with large design effects), design-based variance estimators tend to have degrees of freedom much less than the actual sample size. In another example, design-based regressions using several covariates may have an associated design-based *t*-statistic having a skewed distribution and very low degrees of freedom while the actual nominal sample size is fairly large. Such behavior is typically due to the impact of sampling weights and clustering, which may introduce inefficiencies into the estimation.

To compensate for these design-based inefficiencies and inflexibilities, model-based procedures are often used as an alternative analytic strategy for survey analysis. One such modeling strategy is to use the so-called "random effects" or "multi-level" models. This paper will consider some basic random effects model-based analyses and compare the results to corresponding design-based analyses for situations where design-based methods are appropriate. If the model-based methods perform well in these situations, then such findings may imply the appropriateness of model-based methods in situations when design-based methods have shortcomings. As we are targeting data analysts as our audience, all analyses will be implemented with commonly available software packages.

Comparing a model-based method versus a design-based method using existing NHIS data can not reveal the sampling properties of both methods since the NHIS data are only a realized sample from the underlying population. For an objective comparison we have created an artificial population containing selected geographical, demographic and health variables based on nine years of the NHIS data. This artificial population has structural clustering somewhat consistent with that of the NHIS population, and we have defined a 5-level multi-stage sampling rule that is consistent with the actual NHIS multi-stage sampling for both the all-person per household NHIS sample and the one-adult per NHIS household sample.

To study sampling properties of estimators, a NHIS-type multi-stage sample of about 1000 households are selected using this sampling strategy and then NHIS procedures are implemented on the sample. For analyses, sampling weights are adjusted by post-stratification factors to form final survey weights, and for design-based variance estimation, the multistage sampling is simplified as a "with-replacement sample of 2-clusters per stratum" design (i.e., the ultimate cluster method). Such a simplification is specified on public-use NHIS micro-data and used in many other surveys with complex designs. For this study 1000 independently drawn complex samples are selected from the artificial population, and Monte Carlo methods are used to evaluate sampling distributional properties of survey-based statistics. The details of the pseudo population and the multistage sampling are discussed in Parsons and Parker (2012).

For the present discussion with the emphasis on random effects models, the clustering levels within the population are a key element for the modeling. For the pseudo universe 5 hierarchical clustering levels are defined by Stratum, Primary Sampling Unit (PSU), Secondary Sampling Unit (SSU), Household (HSD) and person. Table 1 shows a decomposition of population variance (the usual S^2 form) into its cluster level components. Stratum and PSU percentages represent the levels of coarse geographical clustering, while SSU percent represents a local geographical clustering. Most of the

variation can be attributed to the HSD and person within HSD. It should be noted that a small magnitude of within-HSD variation is associated with a large intra-HSD correlation and vice versa. This is noticeable for the race/ethnicity and income related variables.

To study properties of statistical analyses based on the complex sampling, we will consider some simple models for body mass index (BMI), treated as a continuous response and smoking status treated as a binary response. This current write-up is based on preliminary findings using the simulated "one-adult-per-household" sample. Model-based results for the "all-person-per-household" sample are anticipated for dissemination in a future article.

2.0 Models, Population Structures, and Data Analysis:

2.1 Underlying population and sampling structures.

Consider a situation where the analyst wishes to determine the impact of gender, race and age upon BMI or smoking status. Either a design- or model- based regression could reasonably specify the fixed effects as

Intercept + Sex (M, F) + Race (White, Black, Hispanic) + Age. (*)

Mathematically this can be expressed as $X\beta$, where X is the matrix of covariates for the sample, and β are the unknown parameters to be estimated.

Now, for a finite population structure the definitions of the population regression parameters, β, on its population typically take the form of a solution to some set of well-defined equations on that population. In a linear regression context, the population parameters are defined $\beta = (X'X)^{-1} X'Y$, where the Y -vector is a complete population variable, and the X matrix consists of columns of complete population covariates. This is in contrast with a super population defined β, which is a fixed vector. (See Binder (1983) for detailed discussion of this approach.) For this research we will consider the population parameter, β, defined by a regression equation to have true value taken as the ordinary least squares solution for continuous Y or the pseudo maximum likelihood solution to that regression equation for binary Y for the equations when applied to the full population.

Under this finite-population definition, design-based regression analyses can be implemented in a well-defined way. Now, in contrast, a model-based regression using the same fixed effects defines a super population structure on the sample data. Modeling requires special consideration of the design features and their incorporation into the model.

An approach frequently taken is to model the random selection processes of the complex sample as random-effects components. For the "one-adult-per-household" sample under consideration, the randomly selected clusters are PSU, SSU, and HSD, but since only one adult is selected per household, only the PSU, SSU and person can be considered as distinguishable random components. If Y represents a normally distributed random variable, then a random intercept model with one random cluster can be expressed as follows:

$Y_{ij} = \mu + x_{ij}\beta + b_i + \varepsilon_{ij}$, $b_i \sim N(0, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, where

Y_{ij} is the observation for subject *j* in cluster *i*,

μ is the intercept,

β is a vector of fixed effects associated with covariates x_{ij} associated with unit *ij*,

b_i is a random effect due to cluster *i*,

ε_{ij} is a residual deviation ('error' in samples), and the random components ($\underline{b}$, $\underline{\varepsilon}$) are mutually independent.

If *Y* represents a binary variable, then a random intercept model with one random cluster can be expressed as follows:

Let b_i be a random effect due to cluster *i*, where $b_i \sim N(0, \tau^2)$, then a Bernoulli response, Y_{ij}, for subject *j* within cluster *i* can be modeled

$P(Y_{ij} = 1 \mid b_i) = \pi_i(b_i)$ where $\pi_i(b_i)$ is defined by

$\log[\, \pi_i(b_i) / (1 - \pi_i(b_i)) \,] = \mu + x_{ij}\beta + b_i$, where the μ, β, *and* x are defined as in the normal model.

These definitions can be expanded in an additive fashion to include additional random effects to form additive random intercept models. More complicated models involving interactions of random effects can be created, but the additive approach is a reasonable starting point.

In this paper we will model the variables BMI and smoking status using the normal and binary structures given above and restricted to additive random effects. The survey weights will not be explicitly used in the model, but as the black households were oversampled and the post-stratification variables were age, race, sex, the modeled fixed effects in equation (*) is assumed to compensate for those weights.

2.2 Analysis procedures:

A design-based analytical approach would specify the design features of survey weights and clustering variables to the designated software in an appropriate way to compute estimated population regression parameters and design-based variance estimates for those parameters. Even though the sampling involves multiple levels of sample selection, only the first random level, PSU, tends to be used in design-based variance estimation. (West (2012), Wolter (2007)). Several examples of sampling properties of design- based regressions with respect to the NHIS pseudo population were presented in Parsons and Parker (2012).

Whether design- or model-based the fixed regression terms, $\mu + x\beta$, will correspond to those expressed in equation (*) above, and the inclusion of random effect terms will complete the model-based specification.

The Sex and Race covariates will be categorical with reference values at Male and White, respectively, and Age will be continuous, but centered at 45 years. Thus, the Intercept represents a White Male of age 45 years.

For equation (*) the population parameters will be more explicitly denoted:
Intercept ~ WM45 (White, Male, age 45) as the reference intercept,
Sex ~ (F-M) (Female – Male) effect controlling for the other covariates,
Race (level 1) ~ (Black – White) effect controlling for the other covariates,
Race (level 2) ~ (Hispanic – White) effect controlling for the other covariates,
Age ~ (Age – 45).

These parameters will be the target quantities of design- or model- based regressions.

2.3 Evaluation Quantities

Let $\hat{\beta}$ and $\widehat{Var}(\hat{\beta})$ represent a generic regression estimator and its variance estimator, defined by either design- or model- methods. Quality measures of an estimator and its estimator of variance are often assessed by the sampling distributional properties of *Bias, Variance, Mean-Squared-Error (MSE), Relative Bias, Coverage and degrees of freedom*. We define these sampling distributional measures as follows:

i. Here, the sampling distribution is defined with respect to repeated survey sampling in the simulation, not to a super population. The E(·) and P(·) operators used below designate the sampling distribution expectation and probability, respectively. It is important to note that while the model-based method is used conditionally given a realization of one sample, the evaluation is based on the totality of all finite-population samples.

ii. β = population target parameter

iii. $E(\hat{\beta})$ = Expected value of $\hat{\beta}$

iv. $Var(\hat{\beta})$ = Variance of $\hat{\beta}$

v. $\widehat{Var}(\hat{\beta})$ = an estimator of $Var(\hat{\beta})$

vi. E($\widehat{Var}(\hat{\beta})$) = Expected value of $\widehat{Var}(\hat{\beta})$

vii. *Bias* ($\hat{\beta}$) = ($E(\hat{\beta}) - \beta$)

viii. MSE ($\hat{\beta}$) = $Var(\hat{\beta})$ + Bias ^{2}($\hat{\beta}$)

ix. Relative Bias of the Variance Estimator:
RBV($\widehat{Var}(\hat{\beta})$) = $E(\widehat{Var}(\hat{\beta}))$ / $Var(\hat{\beta})$

x. Satterthwaite *degrees-of-freedom* of the variance estimator ,
$DF = 2 \cdot [E\left(\widehat{Var}(\hat{\beta})\right)]^2 / Var(\widehat{Var}(\hat{\beta}))$.

xi. *Coverage*: $P(-t_d < (\hat{\beta} - \beta) / \text{Sqrt}(\widehat{Var}(\hat{\beta})) < t_d)$ where t_d is an (1-α/2)-level cutoff for a *t*-distribution having *d* degrees of freedom.

Discussion of the measures defined above:

viii. The *MSE* ($\hat{\beta}$) is a good overall measure to compare to competing estimators. In a survey setting the design-based method will be the standard for comparison. If a model-based method has MSE comparable to the design-based method in many standard situations, then the modeling may imply appropriateness in more aggressive situations where the design-based methods break down.

ix. The target for the *Relative-Bias of the Variance* of $\widehat{Var}(\hat{p})$ is 1.0; values less than unity indicate that the variance estimator is an underestimate on average, and values larger than unity indicate the variance estimator is an overestimate on average. An underestimate raises more concerns as the situation falsely appears to indicate better precision than is warranted.

x. The *Satterthwaite degrees-of-freedom* (*DF*) can be motivated by assuming that the variance estimator has a central Chi-Squared distribution with *DF* degrees of freedom. Under such a condition the relation for *DF* of expression x is exact. In particular, in the design-based approach with two sampled PSUs per stratum, the *DF* can be reduced to "(number of PSU's – the number of strata)" if the PSU sampling within strata is indeed with replacement, and the magnitude of the sampling variance is constant over the different strata, and each stratum component has a Chi-Squared distribution. For the design-based variance in the examples herein, the target *DF* = 14 when all PSUs contain the targeted characteristics of estimation. In the design-based regression setting, the *DF* value is usually less than the nominal, but it is an unknown quantity.

xi. The *coverage* is the proportion of samples for which a nominal size confidence interval captures the true parameter. From the analyst's perspective, the coverage should be the most important measure as it provides a measure of fidelity of nominal coverage to a true coverage. Expression xi can also be expressed in confidence interval form, i.e., the interval defined by $(\hat{\beta} - t_d \cdot \text{Sqrt}(\widehat{Var}(\hat{\beta})) , \hat{\beta} + t_d \cdot \text{Sqrt}(\widehat{Var}(\hat{\beta})))$ is a (1-α/2)·100% confidence interval for β. Skewness and tail weight of the true distribution of the *t*-statistic may result in non-fidelity with the nominal level.

4.0 Findings

Tables 2 and 3 provide the evaluation quantities for design- and model-based methods for response variables BMI and smoking status, respectively, for the regression equation (*).

The evaluation quantities are based on 1000 simulated samples from the "one-adult-per-household" sample. For this study the *R* packages *survey* and *lme4* were used to implement the design- and model-based analyses, respectively. To be succinct in expression, hereupon, the design-based and model-based methods will be denoted by *D* and *M*, respectively.

A "two-PSUs-per-stratum" design with post-stratification weights treated as sampling weights define the sample design structure for the *D* analyses. The design-based approach is denoted by *D* in the tables.

Two model-based approaches are used, one with a random intercept defined by PSU and denoted *M.P* in the tables and another with two additive intercept terms, PSU plus SSU, denoted *M.PS* in the tables. As previously mentioned, it is assumed that the covariates in the model account for much of the survey weighting.

Below are some general observations.

1. For both BMI and smoking status, the *D*- regression estimators show very little bias while the *M* procedures exhibited some larger magnitudes, e.g. BMI bias for Age. However, the Variances for the *M*- regression estimators were almost always smaller. When the bias and Variance are combined by the *MSE* measure, we have (except for one case) that the *D*- and *M*- estimators have *MSE*'s of similar orders of magnitude, but neither estimator shows general superiority. It should be noted that the definitions of the finite population parameters favor the *D* estimators, as the sample weighted regression coefficient estimators are defined to be structurally similar. The *M* estimator is motivated by super population parameters which don't necessarily correspond to finite population parameters. This conceptual difference may help to explain bias differences.

2. In practice survey inference is made using the estimators $\hat{\beta}$ and $\widehat{Var}(\hat{\beta})$ and not the true values. Except for the black minus white parameter, *B-W*, the *RBV* measure of BMI appears to be slightly more conservative for the *M* estimator than for the *D* estimator. For smoking status the *RBV* measures of the *D* and *M* fluctuate between liberal and conservative values, but the *D* method had a greater tendency to be slightly liberal.

3. The Satterthwaite degrees of freedom for the M methods are much larger than those of the D method. This is expected behavior as the *D*-variance estimation methods are somewhat non-parametric. The hypothesized Chi-squared distributions and quadratic form properties seem to be poor approximations as there are several observed increases of variance DF when two additive random effects are in the model compared to having just one random effect.

4. The coverage rates for BMI regression parameter $\hat{\beta}$ appear to be of comparable quality for the *D* and *M* methods. In practice this means that an analyst would tend to make similar inferences about the BMI population using either method. For smoking status the two M-methods at times showed comparable quality to the *D* method except for two noticeable cases; coverage for the parameter β for WM45 using M.PS and coverage for the parameter β for H-W using M.P were quite lower than the *D* counterpart. It should be noted that the M-method for binary variables has more complicated algorithms for estimating parameters than do the normal model algorithms. In the 1000 simulations it was observed that some realizations experienced convergence issues, perhaps generating outliers in the evaluation. Time constraints limited investigation.

4.1 Discussion and Conclusions

While this study is limited in scope, it is the opinion of the authors that neither the *D*- nor the *M* method shows a general superiority or inferiority over the other. As the cases considered were somewhat typical of standard analyses, the observed results give encouragement to using *M*-methods in more aggressive situations. Some outstanding issues that are still to be researched are listed below.

1. Including interactions among covariates and random effects may be needed for better fitting of the models to data. *D*-methods may break down with respect to variance estimation, i.e., over parameterization on limited available data, but M-methods may still be efficient.

2. As of this writing, the algorithm used in the logistic fitting using the R package *lme4* is to be revised. Thus, the number of cases in the current simulation having convergence issues may diminish, if fewer problematic simulation cases result, then the accuracy of the Monte Carlo methods should improve.

3. Nominal sample sizes were used in the evaluation, but with several design factors used as fixed effects. Effective sample sizes are an alternative method to incorporate design features into the analysis.

4. The complete household sample has intra-household correlations that increase the complexity of the M-based approach. The impact of including household random effects needs to be studied.

References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Botman, S., Moore, T., Moriarity, C., and Parsons, V. (2000). Design and Estimation for the National Health Interview Survey, 1995-2004, *Vital and Health Statistics*, 2(130).

Parsons V., Parker J. (2012) Assessing the impact of simplified design assumptions when analyzing data from public-use complex surveys. *Proceedings of the American Statistical Association.* Survey Research Methods Section pp. 4495-4508.

Pfeffermann, D, (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, Vol. 37, No. 2, pp. 115-136.

R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

Wolter, K.M. (2007). Introduction to Variance Estimation, Second Edition, Springer.

West, B.T. (2012) http://www.isr.umich.edu/src/smp/asda/first_stage_ve_new.pdf

Table 1. Between-Cluster-Variation for Select Universe Variables

Universe age 18+ person variables

	Percentage Between-Cluster-Variation				
Variable	Stratum	PSU	SSU	HSD	Within-HSD
Male	0	0	1	25	73
White	18	5	29	42	6
Black	6	6	39	47	3
Hispanic	19	12	22	42	6
Poverty	3	3	24	69	0
Age	1	1	10	68	20
No insurance	1	1	6	76	16
Fair or poor health	1	1	4	67	28
Education	3	4	19	54	20
Smoker	2	1	12	57	28
Body Mass Index (BMI)	0	0	3	62	34

Table 2. BMI Evaluation Measures

Population Regression Parameters				
WM45	H-W	B-W	Age	F-M
26.37	0.93	1.57	2.24	-0.50

Analysis Methods	
D :	Design-based analysis
M.P :	Model with random effect PSU
M.PS:	Model with random effects PSU + SSU

Bias($\hat{\beta}$)

Method	WM45	H-W	B-W	Age	F-M
D	-0.01	0.00	-0.02	0.03	0.00
M.P	-0.10	0.03	0.06	-0.48	0.16
M.PS	-0.10	0.03	0.05	-0.48	0.16

Sqrt(Variance of $\hat{\beta}$)

Method	WM45	H-W	B-W	Age	F-M
D	0.27	0.47	0.41	0.94	0.33
M.P	0.25	0.44	0.39	0.88	0.31
M.PS	0.25	0.44	0.39	0.89	0.31

Sqrt(MSE of $\hat{\beta}$)

Method	WM45	H-W	B-W	Age	F-M
D	0.27	0.47	0.41	0.94	0.33
M.P	0.27	0.44	0.40	1.01	0.35
M.PS	0.27	0.44	0.39	1.01	0.35

Sqrt(Relative Bias of Variance)

Method	WM45	H-W	B-W	Age	F-M
D	0.97	0.98	0.99	0.97	0.98
M.P	1.07	1.02	0.96	1.00	0.99
M.PS	1.08	1.03	0.97	1.00	0.99

Satterthwaite Degrees of Freedom

Method	WM45	H-W	B-W	Age	F-M
D	11	8	8	12	12
M.P	115	118	146	432	695
M.PS	143	125	138	437	694

True Coverage at nominal 95% level

Method	WM45	H-W	B-W	Age	F-M
D	0.94	0.93	0.94	0.93	0.94
M.P	0.95	0.95	0.94	0.91	0.92
M.PS	0.96	0.95	0.94	0.91	0.92

Table 3. Smoking Evaluation Measures

Population Regression Parameters					Analysis Methods
WM45	H-W	B-W	Age	F-M	D : Design-based analysis
					M.P : Model with random effect PSU
-0.86	-0.51	-0.23	-2.25	0.32	M.PS: Model with random effects PSU + SSU

Bias

Method	WM45	H-W	B-W	Age	F-M
D	-0.01	-0.02	0.00	-0.02	0.01
M.P	-0.09	0.10	0.06	0.05	-0.01
M.PS	-0.16	0.01	-0.02	-0.11	-0.04

Sqrt(Variance of $\hat{\beta}$)

Method	WM45	H-W	B-W	Age	F-M
D	0.15	0.29	0.25	0.51	0.17
M.P	0.15	0.27	0.23	0.47	0.15
M.PS	0.16	0.29	0.24	0.50	0.16

Sqrt(MSE)

Method	WM45	H-W	B-W	Age	F-M
D	0.15	0.29	0.25	0.51	0.17
M.P	0.17	0.29	0.24	0.47	0.15
M.PS	0.23	0.29	0.24	0.51	0.16

Sqrt (Relative Bias of Variance)

Method	WM45	H-W	B-W	Age	F-M
D	0.98	0.97	0.97	0.97	1.01
M.P	1.01	0.91	0.85	1.01	1.03
M.PS	0.98	0.94	0.93	1.02	1.02

Satterthwaite Degrees of Freedom

Method	WM45	H-W	B-W	Age	F-M
D	11	9	9	12	12
M.P	56	74	99	293	619
M.PS	64	70	88	225	415

True Coverage at nominal 95% level

Method	WM45	H-W	B-W	Age	F-M
D	0.95	0.95	0.93	0.95	0.95
M.P	0.92	0.90	0.90	0.95	0.96
M.PS	0.82	0.93	0.93	0.95	0.95