Model-based Methods in Analyzing Complex Survey Data: A Case Study with National Health Interview Survey Data[1]

Rong Wei and Van Parsons

National Center for Health Statistics, Centers for Disease Control and Prevention,
3311 Toledo Rd, Room 3222, Hyattsville MD 20782

Abstract
In this study we consider model-based methods that can be used to account for clustering, stratification and weighting effects in complex-survey-design data. Generalized linear mixed effect models were developed based on the adult sample from the public-release of 2007 National Health Interview Survey (NHIS). For this public-release there are only two available levels of clustering, strata and Primary Sampling Units (PSUs) for use in the model-based method. Model-based multilevel variance/covariance structures were estimated using algorithms given in the SAS procedure GLIMMIX. These model-based methods will be compared empirically with the design-based method of the SUDAAN software, as well as with a fixed effect model in the SAS procedure LOGISTIC.

Key Words: general linear mixed models, multilevel modeling, design-based

1. Introduction

The National Center for Health Statistics (NCHS) releases public-use complex-survey data along with guidance for design-based analyses of these data sets, e.g., analytic code is provided for use with SUDAAN® software (Research Triangle Institute (2008)). These design-based methods are considered properly implemented only when all the available design features are considered as input variables to the analysis. The general recommendation is to avoid analyses on a subset of the database unless any lost design information can be retained and incorporated into the analysis. To avoid design structural issues arising from database reduction, it is suggested that the complete data set be used when using standard complex-survey software. However, in some circumstances when large numbers of strata and/or sampling clusters have no valid data or when limited data are linked to other surveys/databases, it may be problematic to use a design-based method that requires complete design structures. In these cases the flexibility of model-based methods might provide an alternative means to obtain information for the objectives of some studies. Among the various sources of complexity in a survey design, only three of them are typically specified in public-use data: strata, first-level-clustering units (PSUs) and survey weights. We will focus on developing model-based procedures for regression-

[1] *The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.*

type analyses that associate health outcomes with possible covariates; our main concern is to account for the three major survey design components.

Our first goal is to develop model-based methods which give results similar to those obtained when using design-based methods on a complete database. For analysts who prefer a design-based paradigm, model-based methods tend to be more acceptable if the results from a model-based analysis are in agreement with the results from a design-based analysis for corresponding analyses of complete data. The targeted audience is the public-use data user wishing to use model-based methods, but while accounting for survey design features. Data from the National Health Interview Survey (NHIS) will be used for this investigation. The work described below should be considered preliminary and exploratory in scope.

2. Data and Methods

2.1 NHIS 2007 adult sample data

For this analysis data from the 2007 NHIS Public-Use Adult file will be used. Information about the NHIS sample can be found at the website: http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm.
Following are important public-use design features available on this data:

1. 23,000 sampled adults of age 18 and older
2. 300 Strata, 600 PSU clusters with 2 PSUs per stratum
3. Geography defined at most by region, North East, South, Midwest and West
4. Survey Weights (WTFA_SA) which include adjustments
5. Poststratification control classes, Sex/Race-Ethnicity/Age
6. Pre-Poststratification weights which include the nonresponse adjustment

The health outcome studied is based on Body Mass Index (BMI), which is a continuous variable, calculated as individual self-reported body weight divided by self-reported height squared in metric units (kg/m^2). Obesity is calculated as a binary variable which has two outcomes: if BMI >= 30, obesity = 1, otherwise, obesity = 0. (Only results for obesity will be included herein.) We will use design-based and model-based methods to analyze obesity.

Explanatory variables from the NHIS adult file include demographic variables:

SEX, AGE, RACE, Hispanic origin (HISP), poverty status (POV),

and health status variables:

Vigorous activity (ACT), Hypertension (HYP), Asthma (ASM), Cholesterol (CHL),
Acid reflux/heartburn (ACI), Headaches (HAC), Alcohol/tobacco (ALC),
Excessive sleepiness (FAT), Depression (DEP), and Anxious (ANX).

2.2 Design Issues

The Public-Use strata and PSUs, while capturing the main components of stratification and clustering, have been subject to masking techniques to reduce geographical disclosure risk. This will result in some structural misspecifications for either design-based or model-based approaches to analyses. For the design-based approaches the clustering has no impact on the first-order estimation (e.g., totals and means), but has an impact on the estimated standard errors.

Most data users treat the adjusted survey weight as a sampling weight (i.e., inverse of probability of selection), and do not attempt to decompose the final weight as part of any analysis. If replicate methods are used, it is possible to replicate many of the adjustments as part of the replicate weight creation process, but currently NCHS does not provide such a set of weights for the NHIS (although, the data user could create such a set from available data and information). Software, like SUDAAN, allows a poststratification to be incorporated (via linearization) into the computation of standard errors for totals and proportions, but not for the more complicated regression-type statistics.

When modeling data, issues on the proper use of the weights and clusters arise. The challenging issue of regression weighting is discussed in Gelman (2007). Figure 1 provides an example of the impact of using the weights or not. In this figure, obesity is computed both weighted and unweighted for sex-race-ethnicity-age groups. Restricting estimation to individual within-group substrata should reduce the impact of the variability of weights. However, we still see major differences between weighted and unweighted obesity. For example, the *non-Hispanic Asian* domains show large disagreements, and the *non-Hispanic other* domains show close agreements between weighted and unweighted estimates. Any attempt at modeling weights will most likely need to include design variables related to geographical sampling and geographical non-response to account for the weighting factors. For public-use data this information is limited.

Clustering was not directly discussed in the Gelman (2007) paper, but some recent discussion appears in Rabe-Hesketh, Skrondal (2006). Figure 2 shows PSU variability within strata for unweighted obesity for white females age 18-64. For this group the weighted and unweighted estimates were very close in magnitude, and only the latter are displayed. As the data within any PSU may be correlated, any model-based approach needs to account for this clustering effect.

2.3 Methods

For this study we look for easily implemented model-based approaches using existing computer software that use "simple" weighting and clustering techniques to provide inference consistent with the design-based approach. We are using the SAS® (2009) and SUDAAN software as presented in Table 1.

Table 1.

Analysis Options	**Accounting for design features**		
	Weighting	Stratification	Clustering
Design-based			
SUDAAN	Yes	Yes	Yes
SAS/SURVEY	Yes	Yes	Yes
Basic SAS Fixed-effects models			
GLM/Logistic	Yes	No[1]	No[1]
SAS Mixed-effects model			
MIXED/GLIMMIX	Yes	Yes[1]	Yes[1]

[1] Strata/PSUs were only used to assess variability and not as fixed predictors of response.

2.3.1 Weighting

To explore methodology for analyzing complex survey data using a model-based regression-type approach, we concentrated on methods that deal with effects due to the design structure, i.e. weighting, stratification and clustering. For public-use data, all levels of sampling and weighting adjustments are summarized as the final NHIS-provided survey weight (See Botman et al., (2000) for a discussion of weighting procedures). Given the challenges of accounting for survey weights in model-based procedures, as a first step we treated the survey weight in a rescaled form $w_{scale,i} \equiv n_{total} w_i/\sum_j w_j$ to accommodate the SAS procedures used. The sum of the $w_{scale,i}$ is n_{total} , the actual total sample size.

As a goal we wanted model-based analyses to yield significance levels of the same order-of-magnitude as a design-based approach and avoid increases in significance levels simply by imposing stronger model-driven assumptions. Preliminary SAS runs with w_{scale} as a weight variable tended to inflate significance levels. To remedy this, we decided to replace n_{total} with an effective sample size. To achieve this we need to reduce the magnitude of the sample by a design-effect factor. To define a design-effect factor to account for weight variability we can use the *coefficient of variation squared* for a set of weights, $\underline{w}$, to define $deff(\underline{w}) \equiv (CV^2(\underline{w}) + 1)$ where all weights used in the analysis are included, see Section 4.4 of Korn and Graubard (1999). An effective-sample-size scaled weight can be defined as $w_{\text{effective scale},i} \equiv w_{scale,i} /deff(\underline{w})$. This factor can be used to reduce the magnitudes of the observed sample size. The use of either scaled weight constrains first-order estimates of ratio-type estimators to be consistent with those produced by using the final survey weight.

It should be noted that $deff(\underline{w})$ adjusts for variability in weights, but not for clustering effects. The factor $deff(\underline{w})$ is defined to be invariant for different analyses on the same database, but clustering is a highly response-variable specific effect. We adjust for clustering specifically through the model.

Usage of these scaled weights can only be assumed to be a “rule of thumb”. For many data users this scaling methodology will be the only option for weighting modifications on “routine” model-based analysis.

2.3.2 Clustering

The SUDAAN data analysis approach is considered “design-based”, in which the so-called between-cluster variance estimator implicitly accounts for intracluster (PSU) correlation. That is, the variance estimator only uses the differences of PSU totals (possibly through Taylor-linearization) within strata, but under suitable sampling assumptions the expectation of the variance estimator is unbiased for the true variance. The between-cluster variance estimator is close to a generalized estimating equation (GEE) approach.

For SAS model-based approaches we have two sampling levels, strata and PSUs. There are numerous options for including the levels into a model, and if included, there is a decision as to define the levels as fixed or random components.

A model-based but *strictly-fixed* effects approach was carried out using SAS/GLM or SAS/LOGISTIC procedure. The model-based mixed effects approach was carried out using the SAS/GLIMMIX procedure. This procedure is able to incorporate multilevel random effects in the estimator, as well as account for the correlation within clusters.

2.3.3 Simple Examples

We consider the NHIS response variable *obesity* along with the covariates provided in Section 2.1 in a logistic regression setting. The computer codes are listed in the box below. The SAS LOGISTIC procedure code is an example of a model ignoring all design variables except the survey weight, and the SUDAAN code represents a typical design-based run. The code from the SAS GLIMMIX procedure represents one attempt to account for the weighting and clustering in the complex design. Options for handling design variables, strata and cluster in model-based analysis by GLIMMIX are:

1. Stratum: included in model as a fixed effect (when number of strata is small) or as a random effect (when the number of strata is large).

2. Cluster: included in model as a first stage random effect, e.g. PSU.

For our data example, a GLIMMIX model specifying both stratum and PSU as random effects required a much longer run time compared to a model with just PSU as single random effect. For comparative runs the GLIMMIX fit statistics and the tests for fixed effects were quite similar, so at this early exploratory stage we a used a one random effect GLIMMIX model, i.e., PSU nested within stratum, to reduce software run times. Also, for comparability, “stratum” was never considered as a fixed effect covariate in any of the models. Furthermore, there seems to be no consensus on denominator degrees of freedom for the *T* and *F* statistics for random effects with unbalanced data. We used 300 degrees of freedom which approximates the *T* and *F* distributions with those of the Z and χ^2 distributions, respectively, and selecting this value is consistent with SUDAAN’s use of the relation *(number of PSUs – number of strata)* as degrees of freedom for variance. SAS has many modeling options that we have not yet fully explored, and the above implementation may possibly be further refined.

Examples of SAS computation codes

1. Simple SAS

```
proc LOGISTIC data= NHIS2007;
class RACE HISP SEX ACT ALC HYP ASM CHL ACI HAC FAT DEP ANX;
model OBESITY =RACE SEX HISP AGE ACT POV ALC HYP ASM CHL ACI HAC
                    FAT DEP ANX;
weight  w_scale  ;    * or w_effective_scale ;
run;
```

2. Model based (many variations)

```
proc GLIMMIX data=NHIS2007;
class strat psu race hisp sex act ALC HYP ASM CHL ACI HAC FAT DEP ANX;
model OBESITY =RACE SEX HISP AGE ACT POV  ALC HYP ASM CHL ACI HAC
                    FAT DEP ANX
                   df =300…300    /dist=binary solution;
random psu_p(strat)   ;
weight w_scale ;  * or w_effective_scale ;
run;
```

3. Design based (SUDAAN)

```
proc RLOGIST data= NHIS2007 design=wr deft4;
nest strat psu;      * provided on Public-Use data file ;
weight wtfa_sa;   * provided on Public-Use data file ;
subgroup race hisp sex act ALC HYP ASM CHL ACI HAC FAT DEP ANX;
levels 3 2 2 2 2 2 2 2 2 2 2 2 2;
model OBESITY =RACE SEX HISP AGE ACT POV ALC HYP ASM CHL ACI HAC
                    FAT DEP ANX;
run;
```

Table 2. |T| -Values of Logistic Regression Parameter Estimates
A Comparison of SUDAAN and SAS PROC LOGISTIC and SAS PROC GLIMMIX
Using Weight Scaling and Clustering Techniques

	SUDAAN RLOGISTIC[1]	SAS LOGISTIC	SAS GLIMMIX[2]	SAS LOGISTIC	SAS GLIMMIX[2]
Weight used →	wtfa_sa	w_{scale}	w_{scale}	$w_{\text{effective\_scale}}$	$w_{\text{effective\_scale}}$
Covariate ↓					
RACE[3]	47.5	152.9	70.7	101.7	49.2
SEX	4.9	5.3	5.2	4.3	4.3
HISPANIC	2.7	3.5	4.1	2.8	3.0
ACT	5.7	7.2	7.3	5.9	5.9
ALC	2.8	3.4	3.2	2.8	2.7
HYP	19.3	23.3	22.9	19.0	18.8
ASM	2.7	3.1	3.1	2.5	2.5
CHL	6.7	8.8	8.8	7.2	7.2
ACI	7.9	9.9	9.6	8.1	8.0
HAC	3.1	3.9	3.5	3.2	3.1
FAT	4.4	5.2	5.2	4.3	4.3
DEP	4.7	5.6	5.6	4.6	4.6
ANX	3.0	3.6	3.4	2.9	2.9
AGE	9.0	10.7	10.7	8.7	8.7
POV	2.4	3.3	2.5	2.7	2.4

[1] SUDAAN uses 300 as denominator degrees of freedom
[2] For GLIMMIX we set all denominator degrees of freedom to 300
[3] No Race effect is expressed by the *F* statistic with 2 numerator degrees of freedom

2.3.3 Evaluation

The comparable fixed effects test statistics of the five modeling runs are presented in Table 2. The SUDAAN run will be treated as the design-based standard for this analysis. In general using the scaled weight, w_{scale} , as the weight option with SAS LOGISTIC or SAS GLIMMIX provides larger $|T|$ and F values, resulting in greater parameter significances than those from a comparable SUDAAN run. SAS LOGISTIC does not account for clustering, and for the two covariates RACE and POV (poverty), which have a tendency for strong geographical clustering, the level of significance of SAS LOGISTIC over SAS GLIMMIX is noticeably larger. Rescaling the weights with $w_{\text{effective\_scale}}$ appears to bring the order of magnitudes of $|T|$ and F closer to those obtained by SUDAAN. The $|T|$ and F statistics for covariates RACE and POV produced by GLIMMIX are now quite consistent in magnitude with those produced by SUDAAN, while LOGISTIC still gives larger values.

3.0 Conclusions

The goal of this research is to find suitable model-based analyses that can be easily implemented with standard software packages and provide results somewhat consistent with those produced by a complex-survey data analysis package. If modeling procedures can be found that work satisfactorily in a somewhat complete data setting, then we feel that the model-based methods can be easily implemented in many partial data settings, e.g., linked data with missing strata and PSUs. While our study is preliminary and quite limited in scope, we feel that the weighting design-effect, *deff*($\underline{w}$), and use of the random effect characterization of the survey clusters, PSUs, show some promise. We are continuing to research along those lines.

References

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

Lehtonen R., Djerf K., Harkanen T. and Laiho, J. (2002). Design-based and model-based methods in analyzing complex health survey data: a case study. *Proceedings of Statistics Canada Symposium 2002*

Lehtonen, R. and Pahkinen, E. J. (2004). *Practical methods for design and analysis of complex surveys, Second Edition:* Wiley

DuMouchel W. H. and Duncan G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, Vol. 78, No. 383 page 535-543.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science,* Vol. 22. No. 2. Page 153-164.

Korn, E., Graubard B. (1999). *Analysis of Health Surveys*: Wiley.

Maas, C.J.M. and Hox J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis,* Vol. 46 page 427-440.

Rabe-Hesketh S., Skrondal A. (2006). Multilevel modelling of complex survey data. *J. R. Statist. Soc. A*, 169, Part 4, pp. 805-827.

Research Triangle Institute (2008). *SUDAAN Language Manual*, Release 10.0. Research Triangle Park, NC: Research Triangle Institute.

SAS Institute Inc. (2009). *SAS 9.2 Help and Documentation*, Cary, NC: SAS Institute Inc.

Figure 1: Comparison of Weighted and Unweighted Obesity Percentages by Sex−Race/Ethnicity−Age; Weighting includes Sample, Non−response, and Post−stratification Factors

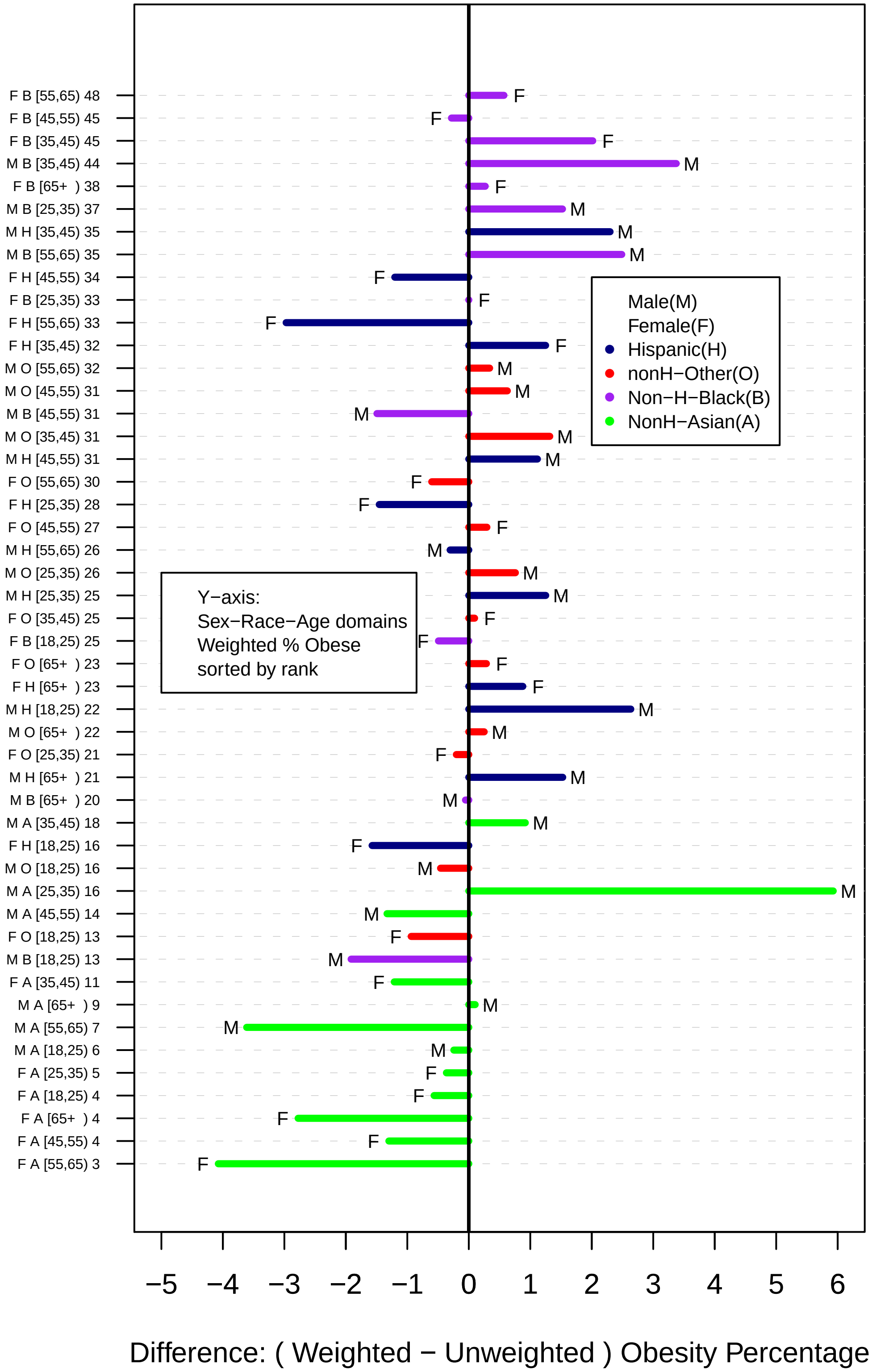
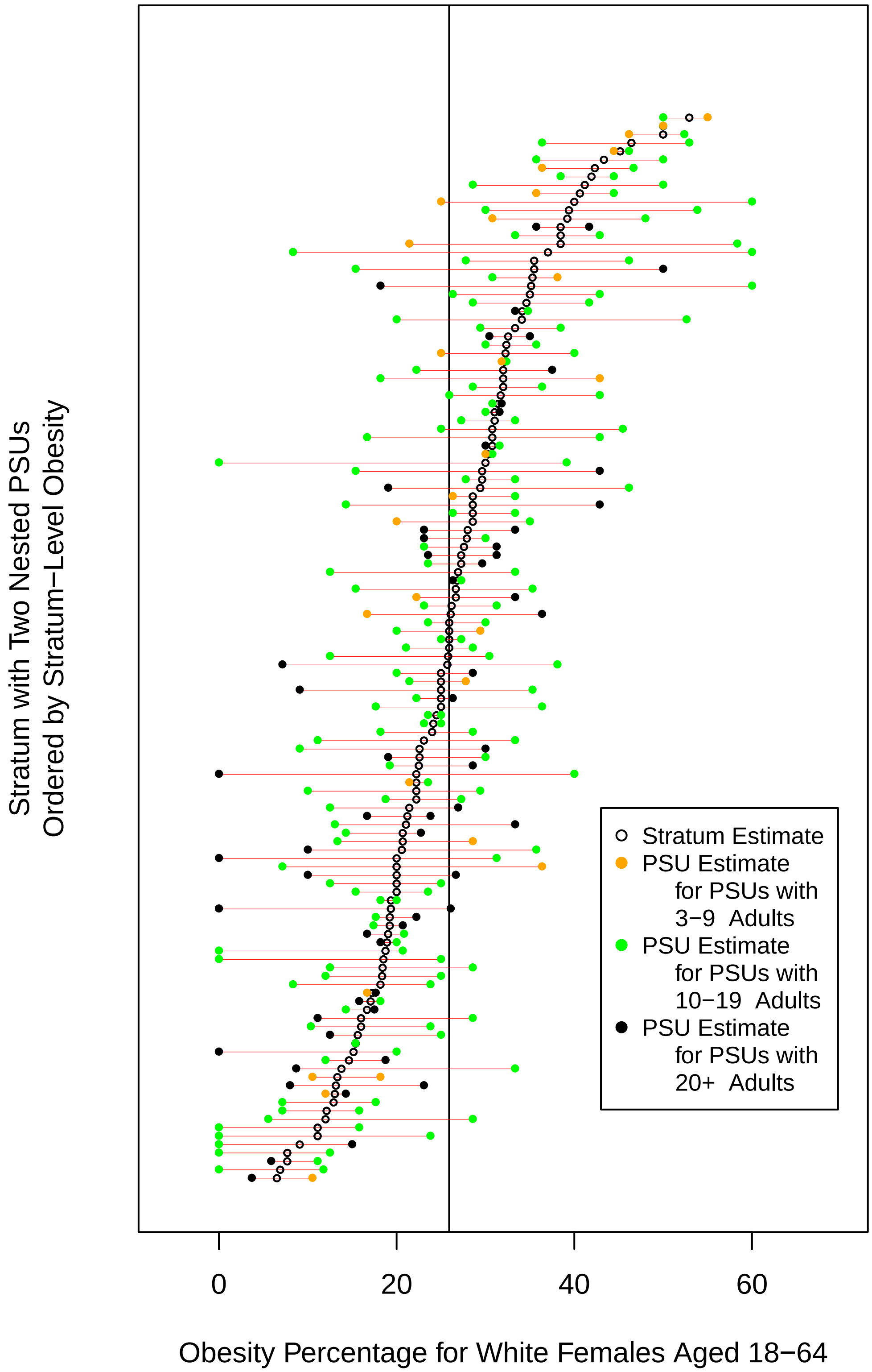


Figure 2: Stratum and Nested PSU Variation Unweighted Estimates; Strata with at least 25 Sampled Adults


Stratum with Two Nested PSUs
Ordered by Stratum−Level Obesity
Obesity Percentage for White Females Aged 18−64
○ Stratum Estimate
● PSU Estimate for PSUs with 3−9 Adults
● PSU Estimate for PSUs with 10−19 Adults
● PSU Estimate for PSUs with 20+ Adults