Inverse Sampling: Investigating a Tool for Model Estimation with Complex Survey Data

Zachary H. Seeskin[1], Edward Mulrow, Josiane Bechara, Qiao Ma
NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603

Abstract
Hinkins et al. (1997) introduced inverse sampling as a way to aid analysts navigating complex sample designs. One intent was to provide users a set of inverse samples that could each be analyzed using methods designed for simple random samples and then combined for inference. These techniques assume one has knowledge of the complex sample design and can properly invert the sample. For public use data, unless inverse samples are provided, a data user would be hard-pressed to create inverse samples based on the complex sample design. Our current research empirically investigates the performance of inverse sampling by comparing the resulting estimates to estimates that use the original sample and incorporate the survey design properly. We study the performance of inverse sampling both when the sampling design is known and when only the survey weights are known and thus only approximate inverse samples can be obtained. The results show that inverse sampling performs well for producing unbiased estimates of model coefficients, but caution is needed for estimating standard errors.

Key Words: Analysis of complex survey data; Resampling; Public use data

1. Introduction

Complex survey designs are tremendously valuable to researchers and data users, allowing data to be collected efficiently and assuring that the data have strong statistical properties to support analyses. Software packages for applying a range of analyses to different complex survey designs are available with common statistical software.

Often, data users are interested in fitting statistical models to survey data, but do not have the resources to use the common statistical packages for estimation. Sometimes, the data user does not have statistical software available to incorporate the survey design for estimating his/her model. It is possible that the statistical model being estimated is sophisticated enough that the methods and/or software for incorporating the survey design for estimation have not been developed. Finally, it may be the case that the user has not been trained in the features of complex survey designs needed to use the software and may prefer a different option.

Hinkins et al. (1997) introduced inverse sampling to support such data users with analyses with complex survey data. Inverse sampling refers to drawing a subsample of a survey sample such that the subsample has the properties of being a simple random sample of the population. Thus, an inverse sample can support analyses assuming that the data are generated by an independent, identically distributed (IID) sampling process.

[1]Address correspondence to Seeskin-Zachary@norc.org.

To address the decreased precision of estimates from using a subsample of the original sample, multiple inverse samples can be drawn, with analyses combined across the inverse samples. Hinkins et al. (1997) demonstrated how to conduct estimation with multiple inverse samples. Using an example for estimating sample means, they show that inverse sampling produces unbiased estimates and the correct variance estimates as long as a sufficient number of inverse samples are used.

Inverse sampling relies on the development of algorithms tailored for specific designs. Further, not every possible complex survey design has had an inverse sampling algorithm developed for it. Thus, a related method for users of complex survey data was applied in Hinkins et al. (2009) called pseudo-inverse sampling. This approach takes subsamples with replacement of complex survey data with subsample selection probabilities proportional to the survey weight. This method can be applied to every complex survey dataset for which weights are available, including public use files which may not provide sample design information beyond the weights.

This paper assesses the value of inverse sampling as well as pseudo-inverse for obtaining analyses with correct statistical inferences. We focus on the use of inverse sampling as a method for model estimation with complex survey data, extending the work of Hinkins et al. (1997). In particular, we focus on two different contexts: when full information about the sample design available and when a user encounters a public use file with limited information about the sample design.

Our studies of linear and logistic regression found that while both proper and pseudo-inverse sampling approaches tended to produce unbiased estimates, neither reliably produced variance estimates that were similar to those obtained from incorporating the sample design in model estimation, nor did the variance estimates consistently over- or underestimate the variance. Thus, inverse sampling or pseudo-inverse sampling can be used when the estimates or model coefficients are of interest. However, further study or refinements are needed before using either inverse sampling or pseudo-inverse sampling for statistical inference.

Section 2 provides further background on inverse sampling and its past uses. Section 3 describes the methods for evaluating the use of inverse sampling for modeling, and Section 4 presents results. Section 5 discusses the findings and suggests some directions for future work.

2. Background

2.1 Inverse Sampling Background

Inverse sampling was introduced by Hinkins et al. (1997) to provide a new tool for users conducting estimation with complex survey data. The authors considered that many common statistical methods are developed with the simplifying assumption that the data are generated by an IID sampling process. The goal of the new method was to allow a user interested in a method relying on this assumption to conduct the analysis on inverse samples, for which the IID assumption would be valid.

For inverse sampling, a specific algorithm must be tailored to the complex sample design. Hinkins et al. (1997) describe the algorithms for inverting a stratified sample, a range of cluster samples, and few kinds of multistage design. A limitation that the paper recognizes is that an inverse sampling algorithm has not been developed for every sample design, and

when the sample design is too complex, inverse sampling may not be possible. The approach for drawing an inverse sample from a stratified sample is presented in Exhibit 1.

Exhibit 1: Inverse Sampling Algorithm for Stratified Sample

1. Set subsample size m to be size of smallest stratum.
2. Determine the number of cases drawn from each stratum by drawing from a hypergeometric distribution:

$$\Pr\left(m_1 = i_1, m_2 = i_2, \ldots, m_h = i_h\right) = \frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\cdots\binom{N_h}{i_h}}{\binom{N}{m}}$$

   where h is number of strata, N is population size, $N_1,...,N_h$ are populations of each stratum, and $m_1,...,m_h$ are the sample sizes to be drawn within each stratum.

3. Draw simple random samples with sample sizes $m_1,...,m_h$ without replacement.

It is straightforward to show that each subsample has probability $1/\binom{N}{m}$ of being drawn.

As for the stratified sample example in Figure 1, the inverse sample size can be much smaller than the sample size, resulting in decreased precision of estimates from using one inverse sample. Thus, Hinkins et al. (1997) propose using resampling, i.e., taking multiple inverse samples with replacement and combining estimates from across the inverse samples.

2.2 Uses of Inverse Sampling

Hinkins et al. (1997) empirically study the use of inverse sampling for estimation from a stratified sample of corporate returns from the Statistics of Income database. They show that using inverse sampling produces unbiased estimates of total corporate assets without using information about the sample design for analysis. They also show that with a sufficient number of inverse samples, the variance of estimates becomes arbitrarily close to the variance of the estimate from stratified sample. With 1,000 inverse samples, the relative increase in variance from using inverse sampling is 3%.

Beyond estimation, inverse sampling has also been used for visualization of regression diagnostics from complex survey data in Hinkins et al. (2009). As many common regression diagnostics are developed for IID data, the paper demonstrates how inverse sampling can be used to examine residuals versus fitted values plots, normal quantile-quantile plots, scale-location plots, and plots of Cook's distances for complex survey data. The paper proposes and demonstrates using a plot corresponding to each inverse sample produced, using an example of a regression with Survey of Consumer Finances public use data.

Because Hinkins et al. (2009) uses a public use dataset without full information about the sample design available to the data user, proper inverse samples per the algorithms developed in Hinkins et al. (1997) cannot be used. Thus, the paper uses pseudo-inverse

sampling, which involves subsampling with replacement from the original sample with probability proportional to the survey weight.

There is limited research on the use of inverse sampling for statistical modeling to date. One investigation was conducted in Nahorniak et al. (2015), which applies pseudo-inverse sampling to estimate models from complex survey data for linear regression, quantile regression, and boosted regression trees. They find for their examples that pseudo-inverse sampling performs well in terms of model coefficient bias and obtaining correct standard errors. However, they do not investigate the proper inverse sampling approach of Hinkins et al. (2009).

3. Methods

3.1 Comparisons between Approaches

To study the performance of inverse sampling and pseudo-inverse sampling, we study linear regression models and logistic regression models, as an example of a non-linear model. We estimate these from two different data sources and compare four kinds of estimates:

A. Estimates based on the original sample that incorporate the survey design, using the *R* survey package. See Lumley (2004) for a description of the package. These results are designated as "Incorporate Design" in this paper. They serve as the benchmark against we compare other methods.

B. Estimates based on the original sample that ignore the sample design, designated as "Ignore Design."

C. When examining stratified samples, estimates conducted via proper inverse sampling using the subsampling approach described in Exhibit 1, designated as "Exact Inv Samp."

D. Estimates conducted via pseudo-inverse sampling, that subsample with replacement with probability proportional to the weight, designated as "Pseudo-Inv Samp."

When incorporating the same design for (A), standard errors are estimated by producing replicate weights using the features of the same design and then applying the replicate weights for estimation.

The estimates from proper inverse sampling and pseudo-inverse sampling are estimated by combining estimates from multiple inverse samples, taking the mean over all inverse samples. Standard errors for all model coefficients are estimated using the bootstrap. Specifically, the standard error is estimated as the standard deviation of parameter estimates from across the inverse samples.

We hypothesized that the estimates from proper inverse sampling will produce unbiased estimates with standard errors similar to those obtained using software to incorporate the sample design. Because pseudo-inverse sampling does not incorporate information about the joint probabilities of sample selection for cases, we hypothesized that pseudo-inverse sampling produces unbiased estimates but does not consistently produce the correct standard errors.

3.2 Data

3.2.1 2000 California Academic Performance Index

The first dataset we examine includes data from 6,194 California public schools as of 2000 with a variety characteristics and a score summarizing students' academic success, called the Academic Performance Index (API). The *R* survey package makes these data available and includes different samples, including the original population. All sample design information is available. We use the stratified sample (sample size 200 with strata of sizes 100, 50 and 50) and the two-stage cluster sample (sample size 126) made available in the package.

We examine two models, a linear model with API (*api00*) as the outcome and a logistic regression model with an indicator for meeting the school-wide growth target (*sch.wide*) as the outcome. Both models use the same independent variables: the percentage of students who are English language learners (*ell*), the percentage eligible for subsidized meals (meals), and the percentage in their first year in the school (*mobility*).

For the stratified sample, we compare proper and pseudo-inverse sampling, while for the two-stage cluster sample we examine only pseudo-inverse sampling. In each case we take 20,000 inverse samples. When conducting proper inverse sampling with the stratified sample, each inverse sample has size 50, equal to the size of the smallest stratum. For the pseudo-inverse samples, each subsample size is set to the effective sample size of the parent sample, using Kish's formula $n_{eff} = \left(\sum_i w_i \right)^2 \Big/ \left(\sum_i w_i^2 \right)$ where w_i is the weight for sample member i. The effective sample sizes of the stratified and cluster samples are 169 and 45 respectively.

3.2.2 2016 General Social Survey

In addition, we evaluate inverse sampling for estimating models with a public use dataset that contains limited information about the sample design available. The 2016 General Social Survey (GSS) included 2,867 respondents from a complex survey design, reporting on their attitudes, behaviors, and attributes. On the public use file, some data are provided to support variance estimation, including the primary sampling units and strata created for variance estimation purposes.

We examine two models, a linear model with number of hours of TV per day (*tvhours*) as the outcome and a logistic regression model with the reporting of being very happy (*veryhappy*) as the outcome. Both models use the same independent variables: respondent age (*age*), parent's education (*paeduc*), and respondent education (*educ*). We only examine pseudo-inverse sampling for the GSS. When conducting pseudo-inverse sampling, we took 30 inverse samples of size 2,867, the same as the sample size

4. Results

The regression results with estimated regression coefficients and standard errors are in Table 1. The estimates are compared visually in Figure 1, followed by the standard errors in Figure 2.

Across both datasets, both linear and logistic regression models, and across different kinds of sample designs, the coefficient estimates from proper inverse sampling and pseudo-inverse sampling compare very well to the benchmark that incorporates the survey design

in estimation. We find in nearly all cases that the difference from the benchmark in coefficient estimates using an inverse sampling approach is less than that from ignoring the sample design in estimation. These results demonstrate, as expected, that incorporating the sample design is critical for eliminating bias in coefficient estimates. In these empirical investigations of linear and logistic regression models, inverse sampling is an appropriate approach for estimating model coefficients

The estimates from either proper or pseudo-inverse sampling are not materially different from each other. From this study, it is inconclusive whether one approach performs better than the other for estimating model coefficients. This indicates that for public data users who have limited information about the sample design available to them, pseudo-inverse sampling may be a good approach when model coefficients are of most interest.

Table 1: Estimated Model Coefficients and Standard Errors from Four Methods for Estimation with Complex Survey Data

	Incorporate Design		***Ignore Design***		***Proper Inverse Sampling***		***Pseudo-Inverse. Sampling***	
Independent Variable	***Est***	***St. Error***	***Est***	***St. Error***	***Est***	***St. Error***	***Est***	***St. Error***
	Estimates for Linear Model of api00 from 2000 CA API Stratified Sample							
Intercept	820.89	10.52	794.98	11.74	820.42	11.00	820.43	12.57
ell	-0.48	0.41	-0.64	0.42	-0.48	0.32	-0.48	0.39
meals	-3.14	0.29	-2.87	0.29	-3.15	0.25	-3.15	0.30
mobility	0.23	0.45	0.02	0.47	0.28	0.48	0.27	0.57
	Estimates for Linear Model of api00 from 2000 CA API Cluster Sample							
Intercept	811.49	30.23	821.45	15.05			809.97	20.16
ell	-2.06	1.38	-1.30	0.72			-2.04	0.93
meals	-1.78	1.08	-2.92	0.42			-1.80	0.62
mobility	0.33	0.61	0.58	0.64			0.46	1.05
	Estimates for Logistic Model of sch.wide from 2000 CA API Stratified Sample							
Intercept	0.836	0.476	0.766	0.409	0.783	0.500	0.769	0.569
ell	-0.002	0.014	-0.004	0.013	-0.003	0.012	-0.003	0.015
meals	-0.003	0.010	-0.003	0.009	-0.003	0.010	-0.003	0.011
mobility	0.061	0.034	0.040	0.024	0.068	0.037	0.070	0.041
	Estimates for Logistic Model of sch.wide from 2000 CA API Cluster Sample							
Intercept	0.821	0.756	0.658	0.506			0.849	0.507
ell	-0.055	0.022	-0.064	0.023			-0.057	0.025
meals	0.029	0.018	0.024	0.014			0.030	0.015
mobility	0.015	0.030	0.057	0.038			0.015	0.024
	Estimates for Linear Model of tvhours from 2016 GSS							
Intercept	3.462	0.613	3.459	0.425			3.447	0.406
age	0.037	0.004	0.039	0.004			0.037	0.004
paeduc	-0.034	0.050	-0.032	0.020			-0.036	0.031
educ	-0.148	0.034	-0.152	0.026			-0.145	0.027
	Estimates for Logistic Model of veryhappy from 2016 GSS							
Intercept	-1.425	0.391	-1.360	0.298			-1.428	0.298
age	0.003	0.003	0.001	0.003			0.003	0.003
paeduc	-0.023	0.019	-0.027	0.014			-0.023	0.014
educ	0.052	0.022	0.053	0.018			0.053	0.019

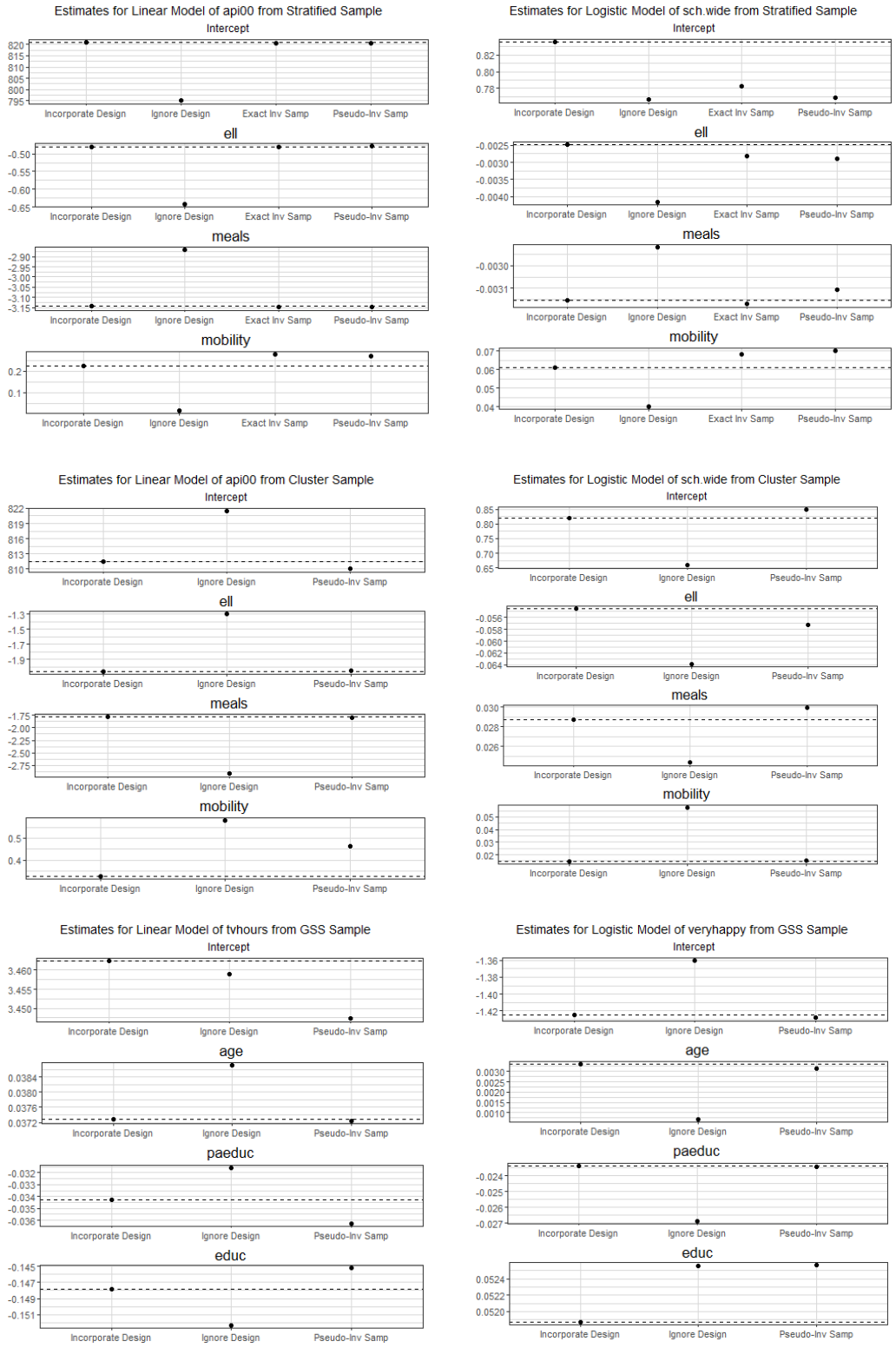

Figure 1: Comparison of model coefficient estimates among incorporating sample design (benchmark), ignoring design, proper inverse sampling, and pseudo-inverse sampling. Examining linear and logistic regression models from 2000 California API (stratified and cluster samples) and 2016 GSS.

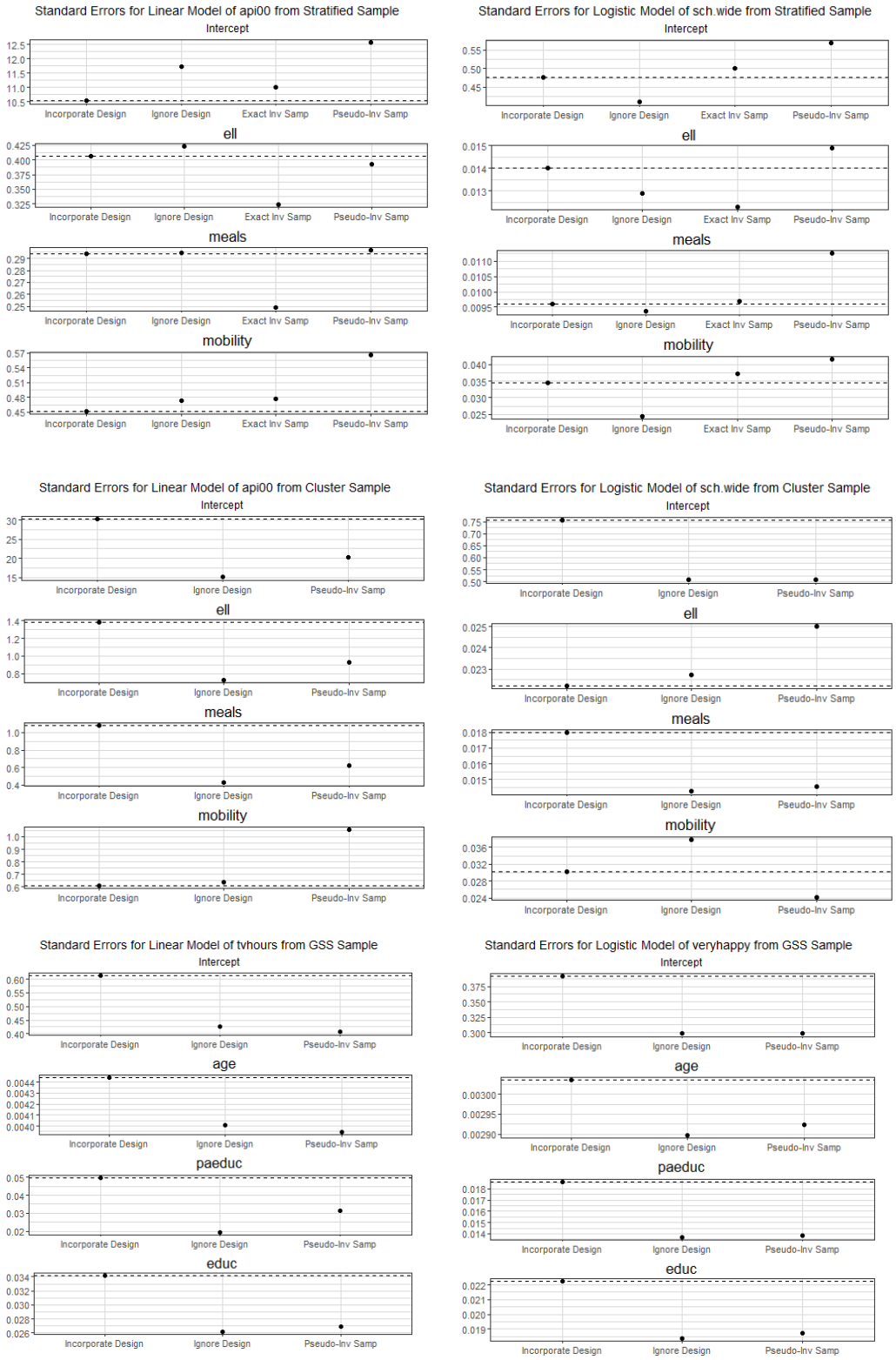

Figure 2: Comparison of model coefficient standard errors among incorporating sample design (benchmark), ignoring design, proper inverse sampling, and pseudo-inverse sampling. Examining linear and logistic regression models from 2000 California API (stratified and cluster samples) and 2016 GSS.

For obtaining standard errors, we do not find a consistent result for how inverse sampling compares with ignoring the sample design, and thus we cannot readily recommend inverse sampling for estimation when standard errors and/or statistical testing are also of interest. Out of eight coefficients for which we examine proper inverse sampling, there are five for which the difference from the benchmark is about as small as or smaller than ignoring the sampling. However, there are instances where using proper inverse sampling led to underestimating the standard error, a finding that surprised us. Particularly for the linear model standard errors for either *ell* or *meals*, using proper inverse sampling yields smaller standard errors and therefore smaller *p*-values for hypothesis testing.

Similarly, the standard error estimates from pseudo-inverse sampling do not consistently perform well when compared with the benchmark. While there are instances where the standard errors compare favorably with proper inverse sampling (e.g, the standard errors for *ell* and *meals* for the stratified sample from the CA Academic Performance Index), there are other instances where the standard errors are either very high or very low compared with the benchmark. This result is not surprising given that pseudo-inverse sampling does not account for joint probabilities of selection, which are critical for variance estimation.

We find that both proper and pseudo-inverse sampling perform well for estimating model coefficients with low bias, but do not perform well consistently for estimating standard errors. The reasons for these findings should be investigated in the future. Users should proceed with caution in using pseudo-inverse sampling for modeling when interested in either the standard errors or hypothesis testing.

5. Discussion

While we caution users as to how to apply inverse sampling, there are some instances where this method can offer strong advantages. First, whether examining linear or nonlinear models, either proper or pseudo-inverse sampling reduces the bias of coefficient estimates and performs much better than ignoring the sampling design. There are instances when the model coefficients are of most interest, and thus inverse sampling may be very useful. Inverse sampling is also helpful when data users cannot readily incorporate the survey design for estimating the model in which they are interested—either because software is not readily available to them or because the model they are estimating may be new or sophisticated enough that the software for their model has not been developed.

Nonetheless, in many contexts, obtaining valid estimates for the standard errors is critical. While variance estimation for complex survey data is an active area of research, the estimates from either proper or pseudo-inverse sampling do not consistently perform well when compared with the benchmark of incorporating the survey design in model estimation. In some cases, the estimated standard errors may be too small. Understanding the reasons why standard errors from proper inverse sampling do not perform well is an area for future research.

Both proper and pseudo-inverse sampling can be computationally intensive, as the methods involve estimating models for each inverse sample. In addition, a proper inverse sampling algorithm has not been developed for every complex survey design. For these additional reasons, data users may face challenges in applying inverse sampling.

In spite of these limitations, we encourage survey statisticians to seek opportunities to understand the needs of their data users and find new ways to meet their needs. There are data users for whom estimating a possibly sophisticated model with complex survey data may prove very challenging, and ignoring the sampling design will yield highly misleading inferences. Future research may prove that in addition to providing survey software, data products may be able to be provided as a solution that would allow estimation of their model more readily. Further, with growing computational power, research on resampling with complex survey data may yield new possibilities for how to analyze such datasets and obtain statistical inferences with good properties.

References

Hinkins, S., Oh, H. L., & Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, *23*, 11-22.

Hinkins, S., Mulrow, E., & Scheuren, F. (2009). Visualization of complex survey data: Regression diagnostics. *2009 Proceedings of the Section on Survey Research Methods*, 2206-2218.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1-19.

Nahorniak, M., Larsen, D. P., Volk, C., & Jordan, C. E. (2015). Using inverse probability bootstrap sampling to eliminate sample induced bias in model based analysis of unequal probability samples. *PLOS ONE*, *10*(6), e0131765.

NORC at the University of Chicago (2017). General Social Surveys 1972-2016: Cumulative Codebook. http://gss.norc.org/documents/codebook/GSS_Codebook_intro.pdf