When or when not to use survey weights? A method for evaluating complex survey design weights

Timothy Raxworthy [1], Yajuan Si [2]

[1] ICF Macro, Inc, 1902 Reston Metro Plaza Reston, VA 20190
[2]Yajuan Si, Institute for Social Research, 426 Thompson St, Ann Arbor, MI 48104

1 Introduction

To begin, it is demonstrated that survey weights can be directly incorporated into the estimation of model coefficients through pseudo maximum likelihood (PML) (Skinner 1989). In this paper there is no direct application of modeling data using the PML method, but for those interested in this topic, we are currently working on a detailed paper using the same data described in this proceeding (Raxworthy and Si 2024). Instead of presenting how the data can be modeled, this paper presents a method for evaluating the survey weights by comparing weighted and unweighted descriptive statistics. We illustrate how survey weights can change estimates using the WHO STEPwise approach to non-communicable disease (NCD) risk factor surveillance (STEPS) survey data. The STEPS surveys are cross sectional multistage stratified surveys, that collect many different risk indicators related to health and aim to provide population level estimates.

2 Weighted Estimation

Following the design of the STEPS survey, unequal probabilities of selection can be accounted for. This is because the sample is a probability sample drawn from a sampling frame. Let $\pi_i = p(i \in s)$ or in other words $\pi_i =$ the probability of selection into the survey sample s for individual i where $i = 1, 2...n$ and n equals the total sample size. Let $y_i = 1$ for person i who has the indicator of interest for our survey and $y_i = 0$ for person i who does not. Once we have defined the probability of selection, adjusted for eligibility, non-response and calibrated the weights to the appropriate population values, we can in theory estimate the population proportion using an expansion estimator otherwise known as a weighted estimator where $w_i = 1/\pi_i$:

$$\hat{y}_s = n^{-1} \sum_{i \in s} \frac{y_i}{\pi_i} \qquad \hat{y}_s = \frac{\sum_{i=1}^{n} w_i * y_i}{\sum_{i=1}^{n} w_i}$$

These expansion estimators are consistent and may be unbiased as the sample size n approaches the total population N of our finite population Q of interest. These expansion estimators are also accompanied by unbiased and consistent methods for estimating the sampling variance for a given estimate using Taylor Series Linearization (TSL) or replication methods such as bootstrap or balanced repeated replication (BRR). To properly compute the sampling variance we need to have codes that correspond to each individual's strata and PSU. After selecting our method for estimating the sampling variance we can define the likelihood function when fitting a logistic regression in the following form:

$$L(y|\boldsymbol{\beta}, \boldsymbol{x}) = \prod_{i=1}^{n} \{\pi(\boldsymbol{x_i})^{y_i} (1 - \pi(\boldsymbol{x_i})\}^{1-y_i} \tag{1}$$

Where $\boldsymbol{x_i}$ is the vector of predictor variables or individual i. The $\pi(\boldsymbol{x_i})$ is the inverse logit link function $\pi(\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i \boldsymbol{\beta}})}$.

This likelihood function does not account for the sample design of how the data were collected. To define the likelihood in a way that does account for the sample design, function (1) is indexed as the product for each stratum, cluster and individual. This has no change on the overall likelihood calculated for the model but it does illustrate the complex sample design used in most surveys. Another missing component to function (1) is that it also does not incorporate sampling weights. To add all of these missing components, equation (1) is rewritten and defined as the pseudo maximum likelihood (PML) function (Heeringa, West, and Berglund 2017):

$$PL(y|\boldsymbol{x},\boldsymbol{\beta},w) = \prod_{h=1}^{H}\prod_{\alpha=1}^{a_h}\prod_{i=1}^{n_a}\left\{\pi(\boldsymbol{x}_{h\alpha i})^{y_{h\alpha i}} * [1-\pi(\boldsymbol{x}_{h\alpha i})]^{1-y_{h\alpha i}}\right\}^{w_{h\alpha i}} \tag{2}$$

For a logistic regression, we estimate the vector of regression coefficients (β) by using the Newton-Raphson algorithm or other optimization method that uses an iterative estimation method for maximizing the likelihood. As shown in the function above we can clearly see how the sampling weights enter into the estimation of the model coefficients when we fit a logistic regression. To compute the variance associated with each coefficient the delta method is used and the estimated variance becomes a combination of the model variance component and the sampling variance component (Binder 1983). These variance estimators are considered to be consistent and unbiased estimators as long as the sampling variance accounts for the loss of precision due to clustering. We will not present this variance estimator but see Binder (1983) for a full presentation on this topic.

3 Bayesian Method for Data Analysis

Survey data collected from complex samples can also be analyzed using Bayesian methods. There are trade offs with going Bayesian as opposed to using a pseudo maximum likelihood function to estimate model parameters but there is no reason to say that one method in totality is more favorable than the other. Using a Bayesian method is particularly well suited for survey data as the sample design usually results in hierarchical data. Estimating the posterior distribution of a multi-level model will increase gains in control over the hyper parameter specification due to the inclusion of priors.

The posterior distribution of the observed data is written as (Hornby et al. 2023):

$$\pi(\theta|y,w) \propto \left[\prod_{i=1}^{n}\pi(y_i|\theta)^{w_i}\right]\pi(\theta) \tag{3}$$

Like the PML function shown earlier, survey weights are included into the estimation of the pseudo-posterior distribution. To demonstrate the hierarchical nature of the posterior distribution, we can alternatively write it as such:

$$\pi(\theta|y,w) \propto \left[\prod_{h=1}^{H}\prod_{\alpha=1}^{a_h}\prod_{i=1}^{n}\pi(y_{h\alpha i}|\theta_{h\alpha i})^{w_{h\alpha i}}\right]\pi(\theta_{h\alpha i}) \tag{4}$$

4 Variable Distribution Comparison

Table 1 in the appendix presents the weighted and unweighted observed distributions for proportion of diabetes, eight age groups, three BMI categories, three education levels and gender for each country. This table also presents each country's sample size and estimated total population size $\hat{N}$. These estimated total population sizes are computed by summing up the survey weights for each country.

There is a large range of total population sizes among the countries included. For example the largest country, Vietnam, has an estimated population size of 55,726,102 where as the smallest country, Kiribati, has 285,325. To put this into other words, one of our groups is approximately 195 times larger than another group. In the case where random intercepts are included in a model specification

and classified by country, we expect that the country specific estimates will be closer to one another than estimates generated by a model that does not include random intercepts classified by country.

When we compare the outcome variable (classified as diabetic or not) between the weighted and the unweighted computations we observe a decrease in the proportion of those with diabetes when we apply the survey weights (aside from El Salvador and Iraq). The largest magnitude of change occurs for Kiribati where we see a decrease of 0.194 in the proportion when we apply the weights. The smallest magnitude of change occurs for Benin with a decrease of 0.004. If we split the countries into two groups where one group are those countries who have a population greater than or equal to 1,000,000 and the other has a population less than 1,000,000, we see that largest magnitude for the smaller population group is 0.194 comparatively to a change of 0.025 for the larger group. From this we can infer that for our outcome variable, the survey weights appear to be more informative for the country's with a smaller population size than those with a larger population size.

For age group, the country observed to have the largest magnitude of change was Tuvala where we observe a decrease of about 15% in the proportion of people in the 55-59 age category and an increase of about 15% in the 50-54 age category when we apply the weights. Tuvala also has a comparatively small country size being at 301,748. We observe a similar pattern to what we observed between our weighted and unweighted values for the proportion of those with diabetes where the smaller countries exhibit a larger change to their distribution when we apply the weights.

When we compare distributions for BMI category the largest change observed is for Kiribati at around 14%, where we observe a decrease in the proportion of BMI category ≥ 30 and an increase in the proportion for the other two categories. For education categories we observe the largest change for Tuvalu where the weights increase the proportion of those who are at an education level of high school or above by about 21%. This massive increase is accompanied by a decrease in the the proportion of those who only received primary school.

The last variable we compared weighted and unweighted estimates for is sex. Again we see a similar trend to what has been observed in our comparisons of the other variables. The largest change in the distribution of individuals is at around a 21% increase to male individuals for Kiribati when we apply the weights. Although Kiribati exhibited the largest change, countries with large population sizes like Iraq also saw a large increase in the distribution of males in their population when we apply the weights of about 10%.

As show in Table 1 there does appear to be evidence that the weights cause quite a large change in the distributions for certain variables across all the countries although this change seems to be greatly magnified for those countries with smaller population sizes. From this evidence we conclude that including the weights into models for estimating diabetes prevalence could cause large changes in estimations of diabetes prevalence when we model its prevalence using these variables comparatively to a model that does not include the weights. This brings us to the next section where we compare models that use weights in their estimation of coefficients versus those that do not and also model predicted values that apply the weights versus those that do not.

5 Conclusion

From observing the descriptive statistics, it is shown that using the survey weights cause a larger change between the unweighted and weighted estimate when using data from countries with with smaller population sizes. From this, it is concluded that the survey weights are increasingly informative when estimating descriptive point estimates for data that includes small groups relative to the total weighted population size. Although this is evaluative true, it does not mean that estimates using survey weights are in anyway more correct. When comparing these diabetes estimates to those published in other sources, there may be cases where the unweighted estimates are closer to estimates found by experts in these fields. Regardless of the situation, this dilemma demonstrates the need to work with experts in the topic of analysis who can properly evaluate survey estimates beyond the statistical component. If the estimates are far from what most other studies are publishing then the weights should be further scrutinized and potentially not used for specific groups in a large data set.

References

Binder, David A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys". In: *International Statistical Review / Revue Internationale de Statistique* 51.3. Publisher: [Wiley, International Statistical Institute (ISI)], pp. 279–292. ISSN: 0306-7734. DOI: 10.2307/1402588. URL: https://www.jstor.org/stable/1402588 (visited on 07/08/2024).

Heeringa, Steven G., Brady West, and Patricia A. Berglund (2017). "Foundations and Techniques for Design-Based Estimation and Inference". In: *Applied Survey Data Analysis*. 2nd ed. New York: Chapman and Hall/CRC, pp. 55–95. ISBN: 978-1-315-15327-8. DOI: 10.1201/9781315153278.

Hornby, Ryan et al. (2023). *csSampling: An R Package for Bayesian Models for Complex Survey Data*. DOI: 10.48550/arXiv.2308.06845. arXiv: 2308.06845[stat]. URL: http://arxiv.org/abs/2308.06845 (visited on 07/19/2024).

Raxworthy, T. and Y. Si (2024). "Fitting multilevel models using STEPS data from multiple countries for estimating health outcomes". [Unpublished manuscript].

Skinner, C. J. (1989). "Domain means, regression and multi-variate analysis". In: *Analysis of Complex Surveys*. Ed. by C. J. Skinner, D. Holt, and T. M. F. Smith. In collab. with C. J. Skinner et al. Wiley, pp. 59–88. ISBN: 978-0-471-92377-0. URL: https://eprints.soton.ac.uk/34696/ (visited on 07/08/2024).

6 Appendix

Table 1: Description of each country data set and comparison of weighted and unweighted variable distributions

Country	Sample Size	$\hat{N}$	Unw. Prop. Diabetes	Prop. Diabetes	Unw. Age 30-34	Age 30-34	Unw. Age 35-39	Age 35-49	Unw. Age 40-44	Age 40-44
Barbados	963	316,226	0.171	0.068	0.091	0.094	0.137	0.123	0.165	0.162
Benin	3492	3,533,685	0.074	0.070	0.216	0.222	0.203	0.215	0.156	0.132
El Salvador	3170	2,333,146	0.136	0.146	0.159	0.153	0.155	0.158	0.142	0.143
Ethiopia	5841	35,098,576	0.038	0.027	0.236	0.224	0.208	0.200	0.155	0.157
Iraq	2815	13,357,414	0.237	0.245	0.173	0.203	0.170	0.151	0.171	0.153
Kenya	3025	16,677,535	0.047	0.039	0.221	0.217	0.199	0.202	0.152	0.179
Kiribati	1474	285,324	0.234	0.040	0.168	0.137	0.192	0.128	0.162	0.084
Mozambique	1390	8,732,614	0.088	0.063	0.222	0.205	0.178	0.189	0.132	0.138
Solomon Islands	1871	418,565	0.097	0.044	0.186	0.167	0.195	0.195	0.165	0.208
Tuvalu	848	301,748	0.172	0.010	0.159	0.244	0.116	0.097	0.113	0.133
Vietnam	3067	55,726,102	0.048	0.037	0.150	0.168	0.147	0.149	0.151	0.157

Country	Unw. Age 45-49	Age 45-49	Unw. Age 50-54	Age 50-54	Unw. Age 55-59	Age 55-59	Unw. Age 60-64	Age 60-64	Unw. Age 65-69	Age 65-69
Barbados	0.138	0.097	0.162	0.217	0.126	0.099	0.110	0.102	0.071	0.103
Benin	0.132	0.177	0.105	0.103	0.071	0.062	0.062	0.050	0.055	0.034
El Salvador	0.143	0.150	0.121	0.120	0.104	0.107	0.990	0.090	0.077	0.079
Ethiopia	0.110	0.124	0.111	0.112	0.068	0.081	0.057	0.052	0.055	0.049
Iraq	0.139	0.145	0.109	0.103	0.083	0.105	0.094	0.092	0.060	0.047
Kenya	0.111	0.117	0.093	0.096	0.086	0.089	0.073	0.055	0.063	0.046
Kiribati	0.147	0.136	0.114	0.193	0.096	0.125	0.069	0.087	0.051	0.109
Mozambique	0.150	0.152	0.140	0.137	0.097	0.096	0.081	0.083	0.00	0.00
Solomon Islands	0.144	0.142	0.113	0.135	0.069	0.064	0.068	0.051	0.059	0.037
Tuvalu	0.128	0.123	0.166	0.314	0.156	0.008	0.104	0.060	0.058	0.021
Vietnam	0.145	0.176	0.132	0.111	0.119	0.106	0.084	0.074	0.071	0.059

Country	Unw. BMI < 25	BMI < 25	Unw. BMI ≥ 25 & < 30	BMI ≥ 25 & < 30	Unw. BMI ≥ 30	BMI ≥ 30
Barbados	0.329	0.372	0.315	0.260	0.356	0.368
Benin	0.650	0.730	0.205	0.172	0.145	0.098
El Salvador	0.276	0.268	0.399	0.407	0.325	0.325
Ethiopia	0.858	0.911	0.091	0.070	0.052	0.018
Iraq	0.176	0.184	0.332	0.354	0.492	0.462
Kenya	0.647	0.668	0.225	0.211	0.128	0.120
Kiribati	0.212	0.281	0.277	0.351	0.511	0.368
Mozambique	0.588	0.653	0.198	0.185	0.214	0.162
Solomon Islands	0.363	0.426	0.343	0.325	0.294	0.294
Tuvalu	0.080	0.100	0.206	0.138	0.713	0.762
Vietnam	0.812	0.816	0.154	0.158	0.035	0.026

Country	Unw. Educat. No School	Educat. No School	Unw. Educat. Primary	Educat. Primary	Unw. Educat. HS or above	Educat. HS or above
Barbados	0.001	0.001	0.132	0.138	0.867	0.861
Benin	0.574	0.590	0.296	0.241	0.130	0.169
El Salvador	0.235	0.193	0.548	0.559	0.217	0.248
Ethiopia	0.631	0.618	0.285	0.317	0.084	0.065
Iraq	0.210	0.200	0.506	0.508	0.284	0.293
Kenya	0.195	0.163	0.494	0.506	0.311	0.331
Kiribati	0.034	0.114	0.474	0.406	0.492	0.480
Mozambique	0.255	0.263	0.514	0.537	0.231	0.200
Solomon Islands	0.098	0.121	0.607	0.605	0.294	0.274
Tuvalu	0.008	0.000	0.529	0.322	0.462	0.678
Vietnam	0.057	0.052	0.284	0.296	0.659	0.653

Country	Unw. Female	Female	Unw. Male	Male
Barbados	0.606	0.409	0.394	0.591
Benin	0.524	0.552	0.476	0.448
El Salvador	0.656	0.558	0.344	0.442
Ethiopia	0.566	0.438	0.434	0.562
Iraq	0.612	0.509	0.388	0.492
Kenya	0.588	0.493	0.412	0.507
Kiribati	0.549	0.326	0.451	0.674
Mozambique	0.601	0.574	0.399	0.426
Solomon Islands	0.549	0.458	0.451	0.542
Tuvalu	0.558	0.679	0.442	0.321
Vietnam	0.557	0.484	0.443	0.516