The Need to Account for Complex Sampling Features when Analyzing Establishment Survey Data: An Illustration using the 2013 Business Research and Development and Innovation Survey (BRDIS)

Survey Methods: Insights from the Field

Brady T. West, Survey Research Center, Institute for Social Research, University of Michigan-Ann Arbor, USA, bwest@umich.edu
Joseph W. Sakshaug, Institute for Employment Research, Germany, joe.sakshaug@iab.de

How to cite this article : West, B. T. & Sakshaug, J.W. (2018). The Need to Account for Complex Sampling Features when Analyzing Establishment Survey Data: An Illustration using the 2013 Business Research and Development and Innovation Survey (BRDIS), Survey Methods: Insights from the Field. Retrieved from https://surveyinsights.org/?p=9435

DOI : 10.13094/SMIF-2018-00001

Abstract : The importance of correctly accounting for complex sampling features when generating finite population inferences based on complex sample survey data sets has now been clearly established in a variety of fields, including those in both statistical and non-statistical domains. Unfortunately, recent studies of analytic error have suggested that many secondary analysts of survey data do not ultimately account for these sampling features when analyzing their data, for a variety of possible reasons (e.g., poor documentation, or a data producer may not provide the information in a public-use data set). The research in this area has focused exclusively on analyses of household survey data, and individual respondents. No research to date has considered how analysts are approaching the data collected in establishment surveys, and whether published articles advancing science based on analyses of establishment behaviors and outcomes are correctly accounting for complex sampling features. This article presents alternative analyses of real data from the 2013 Business Research and Development and Innovation Survey (BRDIS), and shows that a failure to account for the complex design features of the sample underlying these data can lead to substantial differences in inferences about the target population of establishments for the BRDIS.

Introduction

Secondary analyses of complex sample survey data sets available in the public domain further scientific progress in many applied fields. The “complex sampling” that gives rise to these survey data sets often has three key features: 1) stratification of a target population of interest into different divisions of the population that are homogeneous within and heterogeneous between in terms of the measures of interest; 2) multi-stage sampling of clusters of population elements, which are generally either geographic areas or establishments, in an effort to save costs; and 3) unequal probability of selection into the sample for different population elements of interest, including differential probability of response to

the survey for different population subgroups, which gives rise to a need to use weights when computing finite population estimates. A failure of the secondary analyst to correctly account for these sampling features when analyzing the survey data, which recent literature has referred to as *analytic error* (Biemer 2010; Smith 2011; West et al. 2016; West et al. 2017), can lead to substantial biases in population estimates (Heeringa et al. 2017, Section 2.7, Section 7.3) and incorrect population inferences (Wolter 2007, Section 1.1). West et al. (2016) provide detailed illustrations of the implications of these types of analytic errors.

Recent studies in this area have suggested that analytic errors may be quite common in fields that frequently conduct secondary analyses of complex sample survey data sets (West et al. 2016; West et al. 2017). However, all of the work that has been conducted in this area to date has focused on secondary analyses of *household* survey data. Survey data collected from *establishments* are often analyzed to make inferences about business operations in particular countries, and provide important descriptive information for economists and policy makers to research key economic indicators. Given the importance of these surveys to understanding key economic indicators on a global scale, correct analyses of these data are as essential as correct analyses of household survey data to make population inferences about individuals. Unfortunately, none of the research on analytic error to date has considered the magnitude of this problem for establishment survey data.

The objective of this study was to provide insight into the possible magnitude of this problem for establishment surveys using data from a real establishment survey. We begin by examining a small sample of recently-published research articles using establishment survey data to gain an initial understanding of the analytic approaches being used in these articles, and then turn to alternative analyses of real data from the 2013 Business Research and Development and Innovation Survey (BRDIS). We specifically consider the impact of ignoring the complex sampling features of the BRDIS, and implications of this error for the inferences that analysts would make. We conclude with a discussion of our findings and important future directions for analysts and peer reviewers.

A Brief Look into the Analytic Error Problem in Establishment Surveys

To motivate this problem, we reviewed the analytic methods employed in ten highly-cited, peer-reviewed journal articles presenting secondary analyses of establishment survey data. The articles were identified by submitting a search term within Google Scholar. Search terms included “Establishment Survey Analysis” and “Business Survey Analysis”. The top search results with the highest citation counts were screened and deemed “eligible” for inclusion into the sample if they actually presented original analyses of establishment survey data rather than simply referencing analyses conducted in a different article. At the time of screening (January 2017), the ten selected studies had been cited a total of 7,831 times (according to Google Scholar) with a median citation number of 412.5. All articles were published between 1996 and 2009. The impact factors of the peer-reviewed journals in which these articles were published ranged from 0.788 to 5.538, with a median impact factor of 2.855.

After reviewing the ten articles, we coded each article using a set of seven qualitative criteria, including:

1. In what year was the article made available for public viewing?
2. Did the authors mention accounting (in some fashion) for the survey weights?
3. Did the authors mention accounting (in some fashion) for the sample design features (e.g., stratification, cluster sampling, replicate weights) in variance estimation?

4. Did the authors appear to use a *design-based approach* or a *model-based approach* in the analysis (where the "model-based" approaches include those that ignore sampling features entirely in the model specification), per West et al. (2016)?
5. Did the authors appear to use statistical software procedures appropriate for analyzing complex sample survey data?
6. Did the authors use appropriate methods for subpopulation analysis when design-based methods were employed, per Chapter 4 of Heeringa et al. (2017)?
7. How did the authors describe their inferences: with respect to the target population for a given survey (appropriate: e.g., "...an estimated 60% of businesses *in this population* employed non-U.S. citizens in 2000"), or with respect to the sample in hand (inappropriate: e.g., "...60% of establishments *in this sample* employed non-U.S. citizens in 2000")?

We used these criteria to acquire a sense of what analytic approaches (described in the articles) were being employed by the authors of these peer-reviewed studies. Detailed information about the ten articles and our subsequent coding of the articles can be found in the *Supplementary Materials*. We note that our coding of the analytic approaches employed is based only on the available text in the published articles. Word limits may have prevented the authors from fully describing the approaches used, but we feel that clear descriptions of these analytic approaches are essential for readers to understand whether appropriate approaches were employed and whether any population inferences made are legitimate.

Overall, we found that only two out of the ten articles even mentioned using survey weights in the analysis. Both articles employed design-based estimation approaches in their analysis. The other eight articles employed model-based estimation techniques, but did not incorporate the survey weights into those model-based approaches. None of the ten articles reported accounting for sample design features in some fashion when estimating variances. And none of the articles mentioned using survey software procedures that explicitly accounted for sample design features. Six of the ten articles reported conducting subpopulation analyses, but none of them indicated that they used appropriate subpopulation analysis procedures, instead appearing to discard non-subpopulation cases entirely from the analysis. All but two of the articles described their results with respect to the larger target population (as opposed to the sample at hand).

While the results of this small qualitative analysis are certainly not generalizable to all published studies using secondary survey data of establishments, they do suggest that complex sample design features are not always accounted for in analyses of these data. Given their large citation counts and their publication in some very well-known journals (e.g. *American Sociological Review*, *The Quarterly Journal of Economics*, *ILR Review*), it is rather surprising that these articles largely ignored sample design features and relevant survey analysis procedures, or at least did not mention them when describing the analysis performed. We also note in our reading of these articles that only scant details of the sample design were generally provided. In some cases there were clues indicating that stratification and unequal selection probabilities were used in the surveys, but these features were rarely explicitly mentioned in the methods or analysis sections. Other relevant details such as the survey response rate or techniques for handling missing data were also rarely provided. Collectively, the results of this small qualitative study are entirely consistent with larger meta-analyses that have examined this problem in more depth using household survey data (West et al. 2016), and raise concerns about the types of inferences generated in these peer-reviewed studies.

Methods

The 2013 BRDIS: Background

The Business Research and Development and Innovation Survey (BRDIS) is conducted jointly by the U.S. Census Bureau and the U.S. National Science Foundation. This survey program uses the web to annually collect national statistics on research and development (R&D) expenditures from a nationally representative sample of about 45,000 non-farm for-profit businesses, private or public, with five or more employees in the U.S. (see https://www.nsf.gov/statistics/srvyindustry/about/brdis/faq.cfm). The underlying sample design introduces unequal probabilities of selection for establishments of different size via a probability proportionate-to-size sampling procedure, and also features stratification of the target population of establishments based on characteristics related to key survey measures (including R&D activity and NAICS-based industry code; see https://wayback.archive-it.org/5902/20160210141357/http://www.nsf.gov/statistics/srvyindustry/#sd). Final survey weights for the sampled and responding establishments account for unequal probabilities of selection and differential nonresponse rates across different types of establishments. The nonresponse adjustments assume an ignorable missing at random mechanism, meaning that the adjusted weights would result in approximately unbiased estimates assuming that the model underlying the nonresponse adjustments is correct (Heeringa et al. 2017, Section 2.7.3). The respondent data sets also include codes defining the sampling strata to which each responding establishment belonged.

No BRDIS data are ever released to the public, and all restricted-access data files from the BRDIS can only be accessed via Research Data Centers (RDCs) coordinated and administered by staff of the U.S. Census Bureau. The data from the 2013 BRDIS analyzed in this study were accessed via secure servers in the U.S. Census Bureau RDC that has been established at the University of Michigan Institute for Social Research.

Variables of Interest

For purposes of this study, we deliberately chose to focus on a small set of BRDIS variables identified by U.S. Census Bureau staff as being important to BRDIS analysts, and consider some very simple analyses of these variables. These variables included:

- Total worldwide salary of employees (in millions of dollars);
- Total worldwide count of employees (in thousands of persons);
- Total U.S. expenditures on R&D (in thousands of dollars); and
- Total worldwide expenditures on R&D (in thousands of dollars).

We also extracted the nonresponse-adjusted survey weights for responding establishments, in addition to the variable containing the stratum codes.

Analytic Approaches

Our overall analyses focused on national estimation of means for each of the four variables, in addition to standard errors for the estimated means reflecting sampling variance. At the suggestion of BRDIS staff, we also fitted a simple linear regression model predicting domestic expenditures on R&D as a function of total salary expenditures (centered at the overall mean), total worldwide count of employees (also centered at the overall mean), and the interaction between salary expenditures and employee count (to

see if the relationship of, for example, salary expenditures with domestic R&D expenditures varied depending on employee count). More complex models (possibly using transformations of these variables and incorporating additional covariates) could certainly be possible, but we wanted to focus on a very simple example that would still hold some practical meaning for BRDIS analysts.

Following the work of West et al. (2016) in the household survey data context, we considered three general approaches to performing these analyses:

1. Compute all estimates ignoring the design features entirely (no use of weights in estimation, and no use of stratum codes for variance estimation);
2. Compute weighted estimates of all parameters, but ignore the stratum codes when performing variance estimation (using Taylor Series Linearization) and computing 95% confidence intervals for the parameters; and
3. Compute weighted estimates of all parameters, and incorporate the stratum codes into the variance estimation (again using Taylor Series Linearization) for the weighted estimates and computation of the 95% confidence intervals.

We remind the reader that approach (1) above appears to be the most common approach employed by secondary analysts of establishment survey data based on our small sample of highly-cited articles. On the servers secured by the U.S. Census Bureau RDC, we employed procedures in the SAS software for the analyses, including PROC MEANS, PROC REG, PROC SURVEYMEANS, and PROC SURVEYREG. All estimates were rounded according to U.S. Census Bureau disclosure rules.

Results

Estimation of Means

We first consider the national estimates of the means for each variable in addition to their standard errors when following each of the three approaches in Table 1. We note that there were a total of 243 sampling strata according to the 2013 BRDIS sampling design (for Approach 3), and that all estimates in Table 1 (including sample sizes, which reflect counts of responding establishments) are rounded following U.S. Census Bureau disclosure rules.

Table 1: Estimates of means and standard errors of the estimated means for the four BRDIS variables when following the three alternative analytic approaches.

		Approach 1 (no weights, no strata)	Approach 2 (weights, no strata)	Approach 3 (weights, strata)
Variable	Sample Size	Mean (SE)	Mean (SE) [95% CI]	Mean (SE) [95% CI]
Total Salary Expenditures (Millions)	31,000	539.85 (34.42)	20.45 (1.04) [18.42, 22.48]	20.45 (1.00) [18.48, 22.42]

Total Worldwide Employees (Thousands)	31,000	1.28 (0.09)	0.07 (<0.01) [0.06, 0.07]	0.07 (<0.01) [0.06, 0.07]
Total U.S. expenditures on R&D (Thousands)	28,000	11238.10 (941.62)	260.31 (20.24) [220.63, 299.99]	260.31 (17.16) [222.28, 293.33]
Total Worldwide expenditure on R&D (Thousands)	28,000	9867.83 (831.30)	229.63 (17.86) [194.62, 264.65]	229.63 (17.16) [196.00, 263.26]

Considering the results from Table 1, the importance of using the nonresponse-adjusted survey weights in estimation becomes immediately clear. The BRDIS tends to oversample larger and more active establishments, and a failure to use the weights when computing national estimates (which would down-weight larger responding establishments and up-weight smaller respondents) would lead to substantially over-stated estimates of expenditures and establishment characteristics. For example, when ignoring the weights, one would estimate that the average number of employees in the establishments would be approximately 1,280, and when accounting for the weights, one would estimate that the same average employee count is approximately 70. The importance of accounting for the stratified sampling is less obvious, but still apparent. The stratified sampling increases the efficiency of the estimates, and accounting for the sampling stratum codes when estimating variances will reduce the estimated standard errors of the estimates and produce narrower confidence intervals. This is to be expected when the variables of interest vary across the sampling strata (Heeringa et al. 2017).

Another way of interpreting these results is to consider estimates of the mean squared error (MSE) of the estimates computed when using an incorrect approach. Given our earlier review of the ten highly-cited articles, we consider the estimated MSEs of the four estimated means when using the first approach (ignoring the weights and the stratification). We use the weighted estimates based on the approaches employing the weights in estimation as the “true” population means when computing the squared bias portions of the estimated MSEs (where the MSE is equal to the bias squared plus the variance of a given estimate). Using this approach, the estimated MSEs of the four estimated means based on the first (incorrect) approach are approximately 271,000, 1.472, 121,000,000, and 94,000,000, respectively (following the order in Table 1). When comparing these estimated MSEs to those that would only include the variance estimates based on the third (correct) approach, computed by squaring the standard errors in Table 1 (given that these estimates would have no bias in expectation), the differences in the quality of the estimates are substantial.

Estimation of Regression Coefficients

Table 2 presents the estimated coefficients in the linear regression model of interest and their standard errors when following each analytic approach.

Table 2: Estimates of coefficients in the linear regression model for total domestic expenditures on R&D

(in thousands) and standard errors of the estimated coefficients when following the three alternative analytic approaches.

	Approach 1 (no weights, no strata)	Approach 2 (weights, no strata)	Approach 3 (weights, strata)
Predictor	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	3052.35 (855.22)***	254.42 (19.78)***	254.42 (16.53)***
Total Salary Expenditures (Millions, Mean-Centered)	9.48 (0.19)***	9.64 (2.58)***	9.64 (2.49)***
Total Worldwide Employees (Thousands, Mean-Centered)	2951.54 (107.46)***	1967.67 (894.44)*	1967.67 (750.43)**
Interaction	-0.01 (<0.01)***	-0.01 (<0.01)***	-0.01 (<0.01)***
Sample Size	28,000	28,000	28,000
R-squared	0.1883	0.1738	0.1738

*** p < 0.001, ** p < 0.01, * p < 0.05

Table 2 suggests that the final weights for responding establishments were also informative about the coefficients in this simple linear regression model. We note a substantially different estimate of the mean domestic R&D expenditures (the intercept) for an establishment with mean salary expenditures and mean employee count depending on whether the weights were used in estimation, as expected based on Table 1. We also note that the estimated slope for employee count for an establishment with mean total salary expenditures is reduced by about one-third relative to the unweighted estimate. In other words, the positive relationship of employee count with R&D expenditures for establishments with mean total salary expenditures is substantially smaller in the target population (which has a higher relative share of smaller establishments) compared to the sample (which has a higher relative share of *larger* establishments).

Furthermore, a failure to account for the gains in precision of these estimates due to the stratified sampling is apparent, as a marginally significant slope for employee count (again for an establishment with mean total salary expenditures) becomes much more precise when accounting for the stratified sampling (and important at a stricter level of significance). In general, we note the consistent decreases in the standard errors due to the stratified sampling for all of the estimated coefficients. These results point to the importance of accounting for the complex sampling features in multivariable modeling in addition to descriptive estimation, and these results are entirely consistent with those presented by West et al. (2016) in the household survey context.

Finally, we once again found substantially higher estimated MSEs for the estimated intercept and the estimated coefficient for total worldwide employee count when using the first (incorrect) approach. These large differences in the estimated MSEs were again largely arising from the apparent bias of the unweighted estimates.

Discussion

The results of this study clearly indicate that a failure to account for complex sampling features when performing secondary analyses of complex sample establishment survey data can have severe consequences for the estimates computed and the resulting inferences about larger populations of establishments. In the case of the 2013 BRDIS, which involved stratified PPS sampling of establishments, a failure to use the weights in estimation resulted in substantial over-estimates of descriptive features of the target BRDIS population with respect to R&D expenditures and characteristics of the establishments. In addition, a failure to account for the stratified sampling inherent to the BRDIS design resulted in variance estimates for the weighted estimates that were too large, which would lead to overly conservative inferences. These results held up for both descriptive estimation and linear regression modeling, which is entirely consistent with the message presented by West et al. (2016). Simply put, these design features need to be accounted for in secondary analyses of establishment survey data.

Unfortunately, a parallel qualitative study of ten highly-cited peer-reviewed publications presenting analyses of establishment survey data suggested that these design features were rarely accounted for (if at all) in these publications (see the *Supplementary Materials*). If secondary analyses of establishment survey data are published in very influential journals with high impact factors, and important sample design features are not correctly accounted for when analyzing the data and making population inferences, severely misleading conclusions and incorrect knowledge generation may result. Importantly, the authors of these articles may not have had enough space available to describe their analysis approaches in more detail, but we feel that it is essential for secondary analysts of survey data to (at a minimum!) clarify that design features were accounted for correctly in their analyses.

The results of this study point to the importance of ensuring that appropriate analytic guidelines for survey data have been followed correctly during the peer-review process. We would advocate for the establishment of formal rules for any papers presenting secondary analyses of survey data (establishment data or otherwise) by journals accepting these types of secondary analyses. At a bare minimum, these papers should demonstrate that appropriate secondary survey analysis procedures have been employed by the authors, including the use of appropriate weights in estimation and the use of appropriate sample design features in variance estimation. Software for performing these types of analyses is now widespread (Heeringa et al. 2017; West et al., forthcoming), and provided that clear documentation about how the data in a given survey should be analyzed is available from a given data producer, there should be no excuses for failing to account for this information correctly in secondary analyses. These types of analytic errors are a great disservice to the organizations that spend vast sums of money collecting the survey data, and the tax payers who fund these types of data collections more generally.

We conclude by stressing the importance of additional studies like this in other contexts, using more complex analytic approaches. Analytic error appears to be a widespread problem in non-statistical fields, and more general knowledge of this problem would be facilitated by additional studies like this. We hope that these results will encourage data producers to release more informative documentation for secondary analysts, and peer-reviewed journals to be more careful about this issue during the peer-review process.

Appendix

Supplementary materials

SAS code

References

1. Biemer, P.P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, 74(5), 817-848.
2. Heeringa, S.G., West, B.T., and Berglund, P.A. (2017). *Applied Survey Data Analysis, Second Edition*. Boca Raton, FL: Chapman Hall / CRC Press.
3. Smith, T.W. (2011). Refining the Total Survey Error Perspective. *International Journal of Public Opinion Research*, 23(4), 464-484.
4. West, B.T., Sakshaug, J.W., and Aurelien, G.A.S. (2016). How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data? *PLoS ONE*, 11(6), e0158120. doi:10.1371/journal.pone.0158120.
5. West, B.T., Sakshaug, J., and Kim, Y. (2017). Analytic Error as an Important Component of Total Survey Error: Results from a Meta-Analysis. In: Biemer, P.P, Eckman, S., Edwards, B., Lyberg, L., Tucker, C., de Leeuw, E., Kreuter, F., and West, B.T. *Total Survey Error in Practice*. New York, NY: Wiley.
6. West, B.T., Sakshaug, J.W., and Aurelien, G.A.S. Accounting for Complex Sample Design Features in Survey Estimation: A Review of Current Software Tools. *Journal of Official Statistics*, forthcoming.
7. Wolter, K. (2007). *Introduction to Variance Estimation, Second Edition*. New York, NY: Springer.