Environmental Research 120 (2013) 143–144

Contents lists available at SciVerse ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Commentary

Survey data with sampling weights: Is there a “best” approach?

Robert W. Platt [a,b,*], Sam B. Harper [b]

[a] Department of Pediatrics, McGill University, 2300 Tupper St., Montreal, Que., Canada H3H 1P3
[b] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 1020 Pine Ave. West, Montreal, Que., Canada H3A 1A2

ARTICLE INFO

Article history:
Received 10 August 2012
Accepted 15 August 2012
Available online 17 September 2012

Whether and when to use sampling weights in analyses of complex survey data has been the subject of substantial debate in the statistical and epidemiologic literature; an excellent summary of the issues is given by Korn and Graubard (1999). The exchange in this journal (Kim, 2012; Kim and Kim, 2012; Lee and Kim, 2012) presents two different opinions on the issue. Here, we outline some of the challenges and controversies, and provide some guidance for future work.

Complex surveys may sample certain groups or strata at higher rates than others, in order to ensure sufficient precision for group-specific estimates and for bias reduction. Observations are then assigned weights to account for this over-sampling. Additional weighting may also be applied to account for missing data (unit non-response); observations on subjects without missing data are up-weighted in order to represent comparable subjects with missing observations (Brick and Kalton, 1996). An important distinction between the two kinds of weights is that the sampling weights are usually determined by the design, planned in advance, and considered fixed and known, while the missing-data weights are based on a model of the missingness mechanism, i.e., the process that gives rise to missing data.

In a complex survey the weights are designed to ensure that the sample is representative of the whole population when the weights are taken into account. Weighted estimates of population-level parameters (e.g., the mean of a continuous outcome variable) are then unbiased, meaning that if the study were repeated infinitely many times, the average parameter estimate should equal the true population parameter. There is relatively little debate on the need to use weighted analysis for this purpose (Korn and Graubard, 1999); it is essential when the mean outcome differs between sampling groups, or in the complete cases vs. the full data. Weighting may be statistically inefficient when the mean outcome does not differ between sampling groups (meaning that standard errors are large and confidence intervals wider than other unbiased methods), but it nevertheless provides unbiased estimation.

For estimation of associational parameters, the situation is less clear (Korn and Graubard, 1999). There are at least three approaches that can provide unbiased estimation. Weighted analysis is unbiased, but again potentially statistically inefficient (giving rise to wider confidence intervals). Model-based adjustment for sampling (i.e., including variables describing the survey design) is unbiased and can be efficient if models are correctly specified, but this is based on the assumption that the model is correct. Incorrect specification can lead to bias and inefficiency. Unadjusted analyses will be unbiased if there is no effect measure modification by sampling strata; that is, if the association under study is homogeneous in each of the sampling strata. This is a population-level assumption and is not testable.

Kim and Kim (2012) are correct in stating that their inferences are correct for those people who completed data, i.e., internally consistent. However, it is unclear whether such an analysis is generalizable to the entire survey population, and such generalization rests on the untestable assumption that missingness is completely at random. On the other hand, Lee and Kim (2012) correctly point out that weighted analyses are unbiased, and do not rely on such assumptions.

In summary, investigators should approach estimation in complex surveys with care. If one is interested in estimating population parameters, weighted analysis is recommended. For associational parameters, weighted analysis is unbiased, but statistically inefficient. Model-based analysis is unbiased and more efficient, if the modeling assumptions are correct, but these are almost impossible to verify. Finally, unweighted analysis as conducted by Kim and Kim, depends for generalizability on important untestable assumptions but can be much more precise than weighted analyses if these assumptions hold. A conservative approach might involve comparing the various analyses; if weighted and unweighted results are similar, efficiency would favor recommending the unweighted analysis. If they differ, investigation of model-based analyses would allow for better understanding of the impact of the sampling scheme on results.

The most appropriate approach for each specific case depends on the usual statistical trade-off; an unbiased estimate is possible

[*] Correspondence to: Montreal Children's Hospital Research Institute, 4060 Ste Catherine St. West, #205, Westmount, Que., Canada H3Z 2Z3.
Tel.: +1 514 934 1934 × 23288; fax: +1 514 412 4331.
E-mail address: robert.platt@mcgill.ca (R.W. Platt).

http://dx.doi.org/10.1016/j.envres.2012.08.006

with minimal assumptions, while a potentially more efficient estimate depends on validity of assumptions, either about the data or models or both.

References

Brick, J.M., Kalton, G., 1996. Handling missing data in survey research. Stat. Methods Med. Res. 5, 215–238.

Kim, K., 2012. Blood cadmium concentration and lipid profile in Korean adults. Environ. Res. 112, 225–229.

Kim, K., Kim, T.Y., 2012. Reply to commentary "Proper sample-weighted data analysis is required to confirm the association between blood cadmium concentration and lipid profile in Korean adults". Environ. Res. 116, 68.

Korn, E.L., Graubard, B.I., 1999. Analysis of Health Surveys. Wiley-Interscience, New York.

Lee, B.K., Kim, Y., 2012. Proper sample-weighted data analysis is required to confirm the association between blood cadmium concentration and lipid profile in Korean adults: re: Kim, K., 2011. Blood cadmium concentration and lipid profile in Korean adults. Environ. Res. http://dx.doi.org/10.1016/j.envres.2011.12.008. Environ. Res. 115, 66–67; discussion 68.