EFFECT OF REGISTER ERRORS ON QUALITY OF SURVEY ESTIMATES

Ari Veijanen
Statistics Finland, ari.veijanen@stat.fi

Consider the estimation of the means of a survey variable in categories of a register variable. In Finland, unique personal identification codes can be used to obtain the values of register variables for a person in the sample. Registers are usually perceived as reliable but error rates of ten percent or more have been observed (Wallgren and Wallgren, 2014). Effects of misclassification have been studied by Zhang and Fosen (2012), for example. This paper describes how errors in register variables affect statistics as compared with results obtained using true values, which are unknown in practice.

The regional means of a survey variable are often estimated with the help of a model fitted to the sample. What happens when some of the explanatory register variables contain errors? Theory on errors-in-variables and measurement errors shows that the estimate of a slope parameter associated with an explanatory variable containing errors tends asymptotically to zero, as the variance of the error increases. It is not known how errors in auxiliary variables affect calibration, generalized regression estimator (GREG), or empirical best linear unbiased predictor (EBLUP). In a simulation experiment with a synthetic population, a large value was added to an auxiliary variable in the population with probability 0.01. Model-free domain-level calibration was sensitive to this contamination. Model calibration (Wu and Sitter 2001; Lehtonen and Veijanen, 2012), which involves predictions from a model instead of auxiliary variables, was much less sensitive. The mean squared error of GREG and EBLUP increased only slightly due to contamination. GREG was still design unbiased, whereas the design bias of EBLUP was affected by contamination in small domains. All these methods incorporated a mixed model.

Consider the class means of a survey variable Y, when the classification of units is obtained from a register. Because of misclassification, the available classification C' sometimes differs from the unknown true class C. When the sample size increases, even a design unbiased class mean estimator for class c tends to the expectation $E\left(Y \mid C' = c\right)$ given the classification, not to the true expectation $E\left(Y \mid C = c\right)$. The bias does not vanish. Under certain conditions, the bias can be approximated using

$$\left|E\left(Y \mid C' = c\right) - E\left(Y \mid C = c\right)\right| \le \left(1 - P\{C = c \mid C' = c\}\right) \max_i \left|«_i - «_c\right|,$$

where $«_i = E\left(Y \mid C = i\right)$.

Typical sources of errors in a register are coding errors in auxiliary variables and misclassification due to a delayed update.

References

Lehtonen, R. & Veijanen, A. (2012). Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics, 66, 125-133.

Wallgren, A. & Wallgren, B. (2014). Register-based statistics: statistical methods for administrative data. 2nd edition.

Wu, C. & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association, 96, 185–193.

Zhang, L.-C. & Fosen, J. (2012). A modeling approach for uncertainty assessment of register-based small area statistics. Journal of the Indian Society of Agricultural Statistics, 66, 91-104.