Modelling of Survey Data

Olga Vasylyk[1] and Oksana Lagoda[2]

[1]Taras Shevchenko National University of Kyiv, Ukraine, e-mail: ovasylyk@univ.kiev.ua
[2]Kyiv National University of Technologies and Design, Ukraine, e-mail: oksala@ukr.net

Abstract

The fundamentals and assumptions of the model-based approach in sample surveys are presented, such methods as superpopulation modelling and Bayesian modelling are shortly described, and the outline of our study of modelling survey data is given.

Keywords: Bayesian modelling, model-based approach, sample surveys, survey data, superpopulation modelling

1 Introduction

Survey data may be viewed as the outcome of two random processes: the process generating the values in the finite population and the process selecting the sample data from the finite population values. There are three different approaches to sample design and analysis: the design-based approach, the model-based approach, and the design-based model-assisted approach. The advantages and disadvantages of all three approaches have been widely discussed in the literature in recent years.

Since we are interested in using models for survey data, we consider the model-based approach and the design-based model-assisted approach, and investigate how they are applied for solving different problem in sample surveys.

2 Model-based approach in survey sampling: main features

For a population U with N units, let Y = (y_1,…,y_N), where y_i is the set of survey variables for unit i, and let I = (I_1,…,I_N) denote a set of inclusion indicator variables, where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included.

Model-based approach to survey sampling inference requires a model for the survey variables Y, which are treated as random (Little 2004). The model is then used to predict the nonsampled values of the population, and hence finite population quantities Q.(Y)

There are two major variants: superpopulation modeling and Bayesian modeling.

2.1 Superpopulation modeling

Analytic inference from survey data relates to the superpopulation model, but when the sample selection probabilities are correlated with the values of the model response variables even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process.

In superpopulation modeling the N population values of Y are assumed to be a random sample from a "superpopulation" and are assigned a probability distribution $p(Y|\theta)$ indexed by fixed parameters θ. Inferences are based on the joint distribution of Y and I.

2.2 Bayesian modeling

Bayesian modeling requires specification of a prior distribution p(Y) for the population values. Inferences for finite population quantities Q(Y) are based on the posterior predictive distribution $p(Y_{exc}|Y_{inc})$ of the nonsampled values Y_{exc}, given the sampled values Y_{inc}. In this case, model formulations do not involve the distribution for I, basing inferences only on the distribution of Y. This is justified when the sampling mechanism is "non-informative".

Sampling mechanism is said to be non-informative for a variable Y if the distribution of the sampled values of Y and the distribution of the non-sampled values of this variable are the same (Chambers, 2003). Or, in other words, the distribution of I given Y does not depend on the values of Y (Little, 2004).

2.3 Design-based model assisted approach in survey sampling

Design-based model-assisted approach attempts to combine the desirable features of design-based and model-based methods (Särndal *et al.,* 1992*)*.

3 Outline of the study

In our study of modelling survey data, the following main issues were considered:

1. Model-based approach in survey sampling

2. Population models

3. Design-based model assisted approach in survey sampling

4. Model-based and model-assisted estimation for domains and small areas

5. Model-based and model-assisted methods in dealing with nonresponses

6. Weighting and calibration

7. Modelling of complex survey data

8. Models for survey sampling with sensitive characteristics

As a result of the study, a textbook for Master students specializing in Statistics at Ukrainian universities will be prepared.

References

Chambers, R. (2003) *An introduction to model-based survey sampling*. Seminario international de estadistica en eusradi, No.42, 90 p.

Lehtonen, R. (2006) *The Role of Models in Model-Assisted and Model-Dependent Estimation for Domains and Small Areas.* In: Proceedings of the Workshop on Survey Sampling Theory and Methodology (Ventspils, Latvia) pp. 35-44.

Lehtonen, R. (2009) *Estimation for domains and small areas with design-based and model-based methods.* Lectures at the BNU Summer School on Survey Statistics (Kyiv, Ukraine).

Little, R. J. A. (2004) To model or not to model? Competing Modes of Inference for Finite Population Sampling. *The Journal of the Ameriacan Statistical Association*, Vol.99, No. 499. pp.546-556.

Montanari G.E. and Ranalli M.G. *Multiple and ridge model calibration for sample surveys.* Proceedings of the Workshop in Calibration and estimation in surveys, Ottawa, October 2007, Statistics Canada.

D. Pfeffermann (2010) Small Area Estimation: Basic Concepts, Models and Ongoing Research. *The Survey Statistician*, No.62, pp. 26-32.

D. Pfeffermann (2011) Modelling of complex syrvey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology,* Vol. 37, No. 2, pp. 115-136.

Rao, J.N.K. (2005) Interplay between sample survey theory and practice: an appraisal. *Survey Methodology*, Vol. 31, No. 2, pp. 117-138.

Särndal, C.-E., Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Särndal, C.-E., Lundstrom, S. (2005) *Estimation in Surveys with Nonresponse.* Wiley, 212 p.

Valliant, R., Dorfman, A.H. and Royall R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* John. Wiley & Sons, New York.

Vasylyk O., Ianevych T. Using models in sample surveys. *Applied Statistics. Actuarial and Financial Mathematics*, No. 1, pp. 72 – 79, - 2014 (Ukrainian)