Design and Weighting Effects in Small Firm Survey in Korea[1]

Keejae Lee
Department of Information and Statistics, Korea National Open University
169 Dongsung-Dong, Jongro-Ku
Seoul, Korea
kjlee@mail.knou.ac.kr

James M. Lepkowski
The University of Michigan
Ann Arbor, U. S. A.
jimlep@umich.edu

1. Introduction

In many of the national establishment surveys, it is important to make estimates of domain such as industry classification, occupations, gender etc. This requires differential sampling rates so as to obtain adequate sample sizes for small domains. Differential sampling rates require weighting of the sample data. Weighting may also be introduced to compensate for differential non-response. Ignoring the sample weights in an analysis can lead to substantial bias. It is important to separate out the effect of weights on sampling error(or design effect), as this effect tends to inflate sampling error of the estimates. The effect of complex sample design on an estimator can be measured by the design effect, which is the ratio of the variance of the estimator under the complex sample design to the variance calculated as if the sample data came from simple random sampling.

In this paper, we discuss the design and weighting effects on descriptive and analytic statistics and compared several methods for measuring the weighting effect through an empirical study. To compute the standard errors for the labor statistics in small firms, we use the Taylor method applied in the SUDAAN.

2. Results

The sampling design of the 1998 small firm survey in Korea is a stratified one-stage cluster sampling. The sample is composed of 14,942 firms or clusters with a total of 33,116 employees. Due to large number of clusters in the sample, the sample data can be used to evaluate the methods to measure the effect of weights on descriptive and analytic statistics. We conducted an empirical study to investigate the effect on weighting using the 100 EPSEM subsamples selected from the survey data.

Table 1 reports the summary of the three methods to evaluate the inefficiency due to weighting. In summary, Kish's approximation formula to assess the inefficiency loss due to weighting works well.

[1] This study was carried out when the first author was visiting the University of Michigan under the financial support from Korea Science and Engineering Foundation.

Table 1. Inefficiency summary

Variables	Kish's Method	Korn & Graubard's Method	Subsampling method
Monthly salary	45.8%	45.8%	45.5%
ln[monthly salary)		46.8%	45.8%
Normal working hrs		56.4%	47.6%
Extra working hrs		52.0%	46.5%
Age		47.5%	45.8%
Average Inefficiency	45.8%	49.7%	46.2%

We consider survey regression procedure and Generalized Estimation Equation(GEE, Liang and Zeger, 1986) approach to model the salary as a linear function of industry classification, occupation classification, gender, education level, etc. We also consider design and weighting effects on estimation of regression model using the design based approach and the GEE approach.

For the each selected EPSEM subsample, we fit the regression model using the survey regression method and GEE approach with exchangeable correlation structure and calculate the design effect of the regression coefficient estimation. We can evaluate the inefficiency due to weighting through the comparison the design effects of full sample analysis and EPSEM subsamples.

Table 2. Inefficiency summary of regression coefficients estimation

Approaches	Kish's Method	Korn & Graubard's Method		EPSEM Subsample Method	
Design-based approach	45.8%	42.9% [a]	44.7% [b]	43.4% [a]	45.4% [b]
GEE approach with Exchangeable correlation		45.2% [a]	46.6% [b]	43.7% [a]	43.5% [b]

a, b : Average and median of inefficiencies of 30 regression coefficients estimates respectively

In summary, the three methods to evaluate the weighting effects are very similar. The Kish's method seems to be a good approximation of weighting effects in both descriptive statistics and regression analysis.

REFERENCE

1. Korn, E. L. and Graubard, B. I. (1999), *Analysis of Health Surveys*. New York: Wiley.
2. Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear model, Biometriaka 73, 13-22.
3. Shah, B. V., Barnwell, B. G., and Bieler, G. S. (1997). SUDAAN Users Manual, Release 7.5. Research Triangle Park, NC: Research Triangle Institute.

RESUME

In this paper, we conducted an empirical study to investigate the design and weighting effects on descriptive and analytic statistics using the equal weight sub-samples. We compared the regression models using the design based approach and the GEE approach in the view point of the design and weighting effects.