Weighting Imputation for Categorical Data

Liang-Ting Tsai
National Taichung University of Education, Taiwan

Chih-Chien Yang
National Taichung University of Education, Taiwan

Timothy Teo
University of Macau, Macau

INTRODUCTION

LVQ (Learning Vector Quantization) has been used to impute missing group membership and stratum weights in confirmatory factor analysis (CFA) model with continuous indicators (Chen, Tsai, & Yang, 2010; Tsai & Yang, 2012). Currently, categorical questionnaires (e.g., Binary and Likert-type items) are widely used in education, business, economy, and psychology tests as well as international large-scale surveys (e.g., Trend in International Mathematics and Science Study, TIMSS; Progress in International Reading Literacy Study, PIRLS; Program for International Students Assessment, PISA; German Survey of Income and Expenditure, SIE; British Labour Force Survey, LFS). This article aims to adapt the LVQ approach to assess the accuracy of parameters in a CFA model with missing background information in binary and Likert-type questionnaires through a series of simulations.

Questionnaires utilizing categorical and binary items are widely used in business tests and large-scale international surveys. In addition to the responses taken from the items included in the questionnaire, databases used for the analysis of questionnaire results also often provide weighting factors to compensate for non-response bias. This information can be utilized to produce estimates at the level of the population. However, weighting factors in such surveys are unable to consider all the background variables which may affect population level estimates. For example, in the LFS survey, the weight allocated to each individual to better ensure that the respondents were representative of the population was calculated based on age, sex, and region of residence alone (Office for National Statistics, 2011). However, while the researchers conducting the LFS were interested in the relationship between income and economic activity, the survey database did not provide a weighting factor for participant income. Without this weighting factor, a bias would have been introduced on account of the large number of subjects with missing incomes. This type of non-response bias is frequently encountered in the analysis of large-scale questionnaire data, however, to the best of our knowledge no method has been proposed in the literature to account for it. Therefore, to better compensate for this bias and provide more accurate population level estimates, the current study applied the LVQ method to calculate weighing factors for variables of interests.

The concept of sampling weights and the practical applications of survey data have gradually gained importance in advanced statistical models (e.g., CFA, structural equation modeling, multilevel modeling; latent class analysis; latent growth model) (Asparouhov, 2005, 2006; Grilli & Pratesi, 2004; Kaplan & Ferguson, 1999; Patterson, Dayton, & Graubard, 2002; Stapleton, 2002, 2006, 2008; Sonnenschein, Stapleton, & Benson, 2010; Tsai & Yang, 2008; Yang & Tsai, 2006, 2008). To achieve effective results from the analysis of survey data, the analyst needs to adopt proper sampling weights for calculating param-

DOI: 10.4018/978-1-4666-5202-6.ch241

eters in statistical models. However, missing data is a common problem for researchers (Friedman, Huang, Zhang, & Cao, 2012). This is especially the case since the missing data occur in background information, thus also making the sampling weights unrecognizable. Researchers need to appropriately impute the missing information (i.e. Background information and sampling weights) to correctly infer the population characteristics.

To solve the problem of missing group membership and sampling weights, Chen et al., (2010) and Tsai and Yang (2012) find out that the LVQ method can be used to impute missing information. They demonstrate that the LVQ method is as excellent as the weighting-class adjustment (Lohr, 2010) and outperforms the listwise deletion and non-weighted methods of inferring the population parameters of CFA model and identifying measurement invariance (MI) with missing background information. Although LVQ appears to work better than the other methods, evidence only exists for continuous data. A natural extension is to adapt this method for missing data imputation for categorical data. Besides, more simulation studies and sensitivity analysis should be done to obtain solid conclusions to demonstrate that LVQ is a comprehensive method for all sorts of data. It is expected that these four methods will perform similarly in imputing missing data for categorical data as they did for continuous data.

This article is organized as follows. In section 2, the concepts of sampling weights and missing group/background membership are described. In section 3, the detailed procedures of LVQ for imputing missing background information are described. Section 4 describes the experimental design of Monte Carlo study. Results and further research directions are summarized in section 5. Finally, the article concludes with a summary of the results and several thoughts concerning further study.

SAMPLING WEIGHTS AND MISSING GROUP/ BACKGROUND MEMBERSHIP

Researchers have demonstrated the practical applications and increased use of sampling weights in CFA, SEM, and other advanced statistical models when used to analyze survey data (Asparouhov, 2005; Grilli & Pratesi, 2004; Kaplan & Ferguson, 1999; Tsai & Yang, 2008; Yang & Tsai, 2008). Asparouhouv (2005) utilizes factor analysis, latent class analysis, and hierarchical linear models to assess the influences of sampling weights on model parameters. Similarly, Grilli and Pratesi (2004) also imply that sampling weights affect the estimation of multilevel models. Kaplan and Ferguson (1999) establish a CFA model with one factor estimated by six observed variables, and explore the topic of sampling weights using various sample sizes. They conclude that there is a bias in the estimation of factor loadings, which will increase as the differences in the factor loadings between the two research groups continue to increase. Tsai and Yang (2008) and Yang and Tsai (2008) reach similar conclusions as those made by Kaplan and Ferguson (1999).

When sampling weights are missing or undefined, deleting or ignoring them might result in considerable bias of population characteristics (Chen et al., 2010; Tsai & Yang, 2012). However, methods to analyze observations with missing background information and sampling weights have not received sufficient attention. To solve the problem of missing group membership, Chen et al., (2010) and Tsai and Yang (2012) suggest the use of LVQ as a method to allow researchers to categorize information and to make predictions about missing group information. They successfully utilized LVQ to recover missing group membership and weighting information in CFA models. Simultaneously, they also demonstrate