

1 Introduction

There has been a growing interest in recent years in fitting models to data collected from longitudinal surveys that use complex sample designs. This interest reflects expansion in requirements by policy makers and researchers for in-depth studies of social processes over time. Traditionally, the analyses of complex survey samples have been carried out using specialized software packages. More recently, a number of statistical analyses packages, for example SAS and SPSS, have implemented procedures to handle complex survey data appropriately in the case of regression models with continuous and categorical outcome variables. In this guide we describe techniques currently implemented in LISREL 8.7 for analyzing complex surveys data. Research on the longitudinal analysis of complex survey data with LISREL was supported by SBIR grant R43 AA014999-01 from NIAAA to Scientific Software International.

A common theme in substance use research is that data are usually obtained from a multi-stage or so-called complex sampling data. A complex sampling design typically entails stratification, often on the basis of geography; defining meaningful clusters of population elements (PSUs); and one or more stages of subsampling within each PSU. While a complex sample has the advantages of being more economical and practical, guarantees a better representative sample of the population, and does not require a complete sampling frame of the population elements, it is generally less efficient than simple random sampling.

In Chapter 2 some known results are given for the sum of random variables and for multiple linear regression to illustrate adjustments that must be made to accommodate complex sampling properly. In Sections 2.2 and 2.3 we provide a brief summary of important concepts in complex sampling and in Sections 2.4 to 2.6 discuss how these concepts are currently applied to fit regression models to survey data. It is shown that standard software packages for regression analysis allow for a weight variable, but do not yield the correct standard error estimates and measures of fit. In Section 2.7 we present results for the estimation of the covariance matrix of totals and in Section 2.8 an approximate expression for the covariance matrix of a vector of estimated parameters are derived.

The statistical theory and methods for fitting Generalized Linear Models (GLIMs) to simple random sample data are described in a number of textbooks. As pointed out in Chapter 3, inappropriate results are obtained if these methods are applied to complex samples. For quite some time, these methods were extended to include the use of frequency and probability weights in an effort to deal with complex samples. Although this approach yields the appropriate estimates for complex samples, the corresponding standard error estimates are not appropriate. Section 3.2 reviews the options and dialog boxes of the SurveyGLIM menu and SurveyGLIM syntax files are reviewed in Section 3.3. Practical applications are provided in Section 3.4 to illustrate the use of GLIMs for count, continuous, binary, ordinal and nominal response variables. In Section 3.5, the results of the SurveyGLIM module are assessed by means of a simulation study and numerical comparisons with other software. The GLIM statistical theory for analyzing complex survey data is outlined in Section 3.6.

In the social sciences, and increasingly in biomedical and public health research, latent variable (LV) models have become an indispensable statistical tool. A LV is a statistical device to summarize the information in a collection of correlated response variables, thus reducing them to a single new measure. In alcohol abuse studies, for example, latent variables could become a major focus of attention. It is the complexity of attitudes and traits underlying the alcoholism syndrome that is of greatest concern, rather than any specific behavior. As an example, questionnaire items are frequently constructed to deal with the functioning of the subjects in a particular domain. Subsets of these items are often correlated. This implies that the subset reflects a common theme. For example, a possible LV would be Tendency to Use Alcohol. Tendency to Use is a kind of unmeasurable propensity that is more than the combination of these items. The higher the individual Tendency LV score is, the more likely that the person will endorse questionnaire items regarding use and abuse of alcohol. Although structural equation modeling allows for a tremendous flexibility in modeling error structures and for dealing with latent variables, it is in general not straightforward to analyze nested data structures with it. This, on the other hand, is a strong point of multilevel modeling which is also more flexible than structural equation modeling when repeated measurement occasions vary between individuals.

Multilevel models are particularly useful in the modeling of data from complex surveys. Cluster or multi-stage samples designs are frequently used for populations with an inherent hierarchical structure. Ignoring the hierarchical structure of data has serious implications. The use of alternatives such as aggregation and disaggregation of information to another level can induce an increase in collinearity among predictors and large or biased standard errors for the estimates. In order to address concerns regarding the appropriate analyses of survey data, the LISREL 8.7 multilevel module features an option for users to include design weights on levels 1, 2 or 3 of the hierarchy as described in Chapter 4. Section 4.2 gives an overview of the graphical users interface (GUI) for the linear multilevel modeling module. Section 4.3 contains the multilevel syntax that is generated via the dialog boxes. For advanced users, there are additional syntax specifications presently not available via the interface dialog boxes. Practical applications of level 3 models with design weights on the different levels of the hierarchy are given in Section 4.4. In Section 4.5 we provide evaluation and simulation studies. Section 4.6 describes the general weighting strategy of Pfeffermann et al. (1997), followed by a more rigorous theoretical treatment of the implementation of weights, the calculation of robust standard error estimators and the use of fit statistics.

The single most important feature of the LISREL program is its facility to deal with a wide class of models for the analysis of latent variables. Because the whole framework of the LISREL model is based on relationships among LVs, this aspect is discussed in Section 5.1. The general form of the LISREL mode has proven to be so rich that it can handle a large variety of problems. Section 5.2 describes how to draw a path diagram and create syntax using the graphical user's interface of LISREL. An overview of the SIMPLIS syntax, which is used to specify LISREL models, is given in Section 5.3. Illustrative examples are given in Section 5.4. In Section 5.5, a simulation study and empirical comparisons are used to assess the results produced by LISREL in the case of complex

survey data. An overview of the statistical theory implemented in LISREL for the analysis of complex survey data concludes this chapter.