Journal of Statistical Software

April 2018, Volume 84, Book Review 2. *doi: 10.18637/jss.v084.b02*

Reviewer: Abdolvahab Khademi
University of Massachusetts

Complex Survey Data Analysis with SAS

Taylor H. Lewis
Chapman & Hall/CRC, Boca Raton, 2017.
ISBN 9781498776776. xiii+326 pp. USD 89.95 (H).
https://www.crcpress.com/9781498776776

Surveys are commonly administered by international centers, government agencies, research institutions, and businesses to collect and analyze data that reveal population level characteristics. Because the data are collected from samples that may have structural heterogeneity (such as clustering or stratification), analysis of survey data needs a treatment different from the statistical methods ordinarily used in observational and experimental studies. In addition, a different set of computation techniques and software are used for estimation.

One very common statistical software used for survey data analysis and data management is SAS because of its expandability, reliability, educational support, and thorough documentation. There are numerous resources on learning SAS for different data analysis applications, including survey data analysis. An introductory and practical book for this purpose is *Complex Survey Data Analysis with SAS*.

The book is organized in eleven chapters, which progress from basic to advanced and computer intensive techniques. The author states the main aim of the book as the demonstration of survey data analysis techniques for practitioners and researchers who rely mainly or solely on the SAS statistical environment. The book assumes that the readers have a working knowledge of and experience in SAS programming and an intermediate background in statistics and probability.

Definitions and fundamental concepts used in complex survey literature are introduced in Chapter 1, *Features and Examples of Complex Surveys*. The author defines and exemplifies basic concepts such as survey, census, target population, sampling frame, sampling units, and population units. Where confusion may arise or distinctions are blurry (such as the difference between target population and survey population), the author uses clear elaboration, examples, and sometimes graphics, for clarification. SAS commands specific to survey data analysis are briefly reviewed. Next, the author describes the characteristics of complex survey (finite population, stratification, clustering, and unequal weights) in detail and with examples. Throughout, the author contrasts SAS commands designed for survey analysis with the ones commonly used for non-survey data. This chapter ends with three real world survey examples which include complex features explained in the chapter.

In Chapter 2, *Drawing Random Samples Using PROC SURVEYSELECT*, different types of sampling in complex survey analysis are presented, including simple random sampling, systematic sampling, probability proportional to size sampling, stratified sampling, and cluster sampling (and combinations of these types). All sampling methods are illustrated with SAS code, tables, and graphs.

Summary statistics commonly used in survey data analyses are presented in Chapter 3, *Analyzing Continuous Variables Using PROC SURVEYMEANS*. The authors demonstrate how to use the SAS `PROC SURVEYMEANS` command and options to estimate totals, means, ratios, and quantiles in survey data using real data sets. The definitions and mathematical formulation of these estimators are presented and demonstrated using SAS code and output. The author walks through the code and the mathematical formulae to ensure complete understanding.

Complex surveys with categorical data are treated in Chapter 4, *Analyzing Categorical Variables Using PROC SURVEYFREQ*. The author starts out contrasting `PROC SURVEYFREQ` with `PROC SURVEYMEANS` on univariate categorical data, demonstrating the differences and similarities with SAS code and output. In addition to point estimates and interval estimates, the chapter presents inferential tests for categorical data, such as goodness-of-fit tests, tests of association for 2×2 tables, risk statistics, likelihood ratio test, and $n \times n$ tables.

Linear regression on complex survey data is taken up in Chapter 5, *Fitting Linear Regression Models Using PROC SURVEYREG*. The author starts out with introducing the general theory of linear regression, its assumptions, matrix formulation, and hypothesis testing. Later on, the model is respecified for complex survey data and demonstrated through examples, SAS code, and output. Both continuous and categorical explanatory variables are presented. Regression analysis with discrete dependent variables is presented in Chapter 6, *Fitting Logistic Regression Models Using PROC SURVEYLOGISTIC*. The author introduces the application, theory, and interpretation of logistic regression using real data. Once the general theory of logistic regression is outlined, the author presents the method for complex survey data with relevant adjustments. In the end, ordered and unordered multinomial logistic regression is presented.

Statistical methods for time-to-event data are presented in Chapter 7, *Survival Analysis with Complex Survey Data*. The author starts out with defining the terminology for survival analysis in different fields (e.g., biostatistics, engineering, operation research) and then elaborates on the foundations of survival analysis issues, such as data collection designs, censoring, types of survival analyses (parametric, nonparametric, semiparametric, and discrete-time types). Once the reader is familiarized with the general foundation and theory of survival analysis, the method is explained, with example and SAS code, for data collected from complex surveys. The main emphasis in this chapter is the use of Cox proportional hazards regression approach using `PROC SURVEYPHREG` method.

Survey researchers may be interested in a subset of the collected sample at hand. It is tempting to simply subset or filter the larger data set according to some criteria. However, the appropriate method could be more complicated. This issue is taken up in Chapter 8, *Domain Estimation*, where the focus of data analysis is on a subset or a domain of a subpopulation. The author demonstrates an example data set where domain estimation is appropriate. To further emphasize the correct method for subsetting (using proper subsets, domain, and simple filtering), the author analyzes the example data using both methods and then highlights the differences. The author introduces the `DOMAIN` option in the SAS code. However, if the `DOMAIN` option is not available for a specific estimation (such as quantile estimation),

the author recommends a domain-specific weights approach. In the rest of the chapter, the author revisits previous statistical methods (e.g., regression) but now in the context of domain estimation.

Taylor series linearization (TSL) is the default variance estimation method in the `SURVEY` family of procedures in the SAS software. In Chapter 9, *Replication Techniques for Variance Estimation*, variance estimation methods based on repeated resampling are introduced. TSL was not discussed in previous chapters, so before introducing the replication techniques the author presents details about the mathematics and the computation algorithm of TSL. The replication techniques include balanced repeated replication (and Fay's variant), the jacknife method, and the bootstrap replication technique (nonparametric bootstrap). This chapter also demonstrates how replication techniques can be generalized to multivariate statistics, such as a vector of regression coefficients.

Treatment of missing data is discussed in the last two chapters. The author introduces two approaches with regard to nonresponse: in the first approach, the estimation method is adjusted and the data remain intact, and in the second approach the estimation method remains intact but the data is manipulated through statistical recovery of missing data. The first approach is presented in Chapter 10, *Weight Adjustment Methods*, which begins with caveats against traditional listwise deletion of population units with missing data (causing nonresponse error). Nonresponse error and bias are defined and contrasted to demonstrate the effect of listwise deletion on estimators. In addition, survey nonresponse is viewed from deterministic and stochastic perspectives. Next, Little and Rubin (2002)'s renowned classification of missing data is presented. In the rest of the chapter, the author presents four basic weight inflation methods to compensate for missing data, including adjustment cell method, propensity cell method, poststratification, and raking.

The second approach to the treatment of missing data is presented in Chapter 11, *Imputation Methods*. The author defines the deeper goal of data imputation as "to exploit the relationships between the auxiliary variables and the observed data to recapture a portion of the uncertainty caused by the missing data" (p. 275). This chapter begins with a set of definitions and classification of imputations, such as deterministic vs. stochastic, explicit vs. implicit vs. semiparametric, and their cons and pros. Next, application of multiple imputation using SAS code is demonstrated using implicit, explicit, and semiparametric models and various types of variables and statistical methods. In the multivariate section, methods for monotone missingness patterns and arbitrary missingness patterns are presented. Once multiple imputations (with recommended $M = 5$) have been discussed, techniques to reduce multiple estimates to a single estimate are presented in the rest of the chapter using Rubin (1987)'s fundamental combination rules. In the last section of the chapter, the author provides recommendations for applying multiple imputation techniques to complex survey data.

Complex Survey Data Analysis with SAS is a very clear and concise reference for practitioners, students, and researchers who are interested in learning how to analyze data from complex surveys using the SAS statistical environment. The prominent feature of the text is its very clear exposition of concepts in survey statistics combined with implementation code. The author uses clear language, intuitive graphics, contrasts, and real examples to achieve this goal. Although the book is posed primarily as a handbook for SAS (as the titles of the chapters suggest), it nevertheless presents the concepts so clearly that it can also be regarded as an introduction to complex survey analysis.

Another positive aspect of the text is that each chapter is dedicated to particular statistics or a statistical method, with the necessary programming code, output, and interpretation. Therefore, for guidance and help the reader may refer to a particular chapter without having to go through other chapters. The topics selected for the text cover the main estimators and methods practitioners and researchers use routinely in the analysis of complex survey data.

In addition, SAS code to demonstrate analysis of complex survey data is fully reproduced and clearly annotated in the text together with the output. The author makes a fascinating job in clearly walking the reader through the code and interpreting the results. This feature makes the book an indispensable resource for self-learners and practitioners who need a handy reference for using SAS in complex survey analysis.

As stated by the author, the book does not delve into the depths of complex survey theory. However, where a deeper discussion of the concepts would be helpful, the author presents valuable references for the interested readers.

Overall, this is a well-structured and practical desk-side reference for students, practitioners, and self-learners who are interested in performing different data analyses on complex survey data using the SAS statistical software.

References

Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data*. 2nd edition. John Wiley & Sons, New York. doi:10.1002/9781119013563.

Rubin DB (1987). *Multiple Imputations for Nonresponse in Surveys*. John Wiley & Sons, New York. doi:10.1002/9780470316696.

Reviewer:

Abdolvahab Khademi
University of Massachusetts
Department of Mathematics and Statistics
Amherst MA 01002, United States of America
E-mail: khademi@math.umass.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
April 2018, Volume 84, Book Review 2 *Published:* 2018-04-18
doi:10.18637/jss.v084.b02