Estimation in complex sample design with different statistical software packages

Outi Ahti-Miettinen[1]

[1] Statistics Finland
e-mail: outi.ahti-miettinen@stat.fi

Abstract

Limited resources in data collection and requirements of subpopulation analyses lead up to use of complex sample designs. Clustering, stratification and multi-stage sampling often produce data that is inconsistent with assumption of independent observations. This complicates especially the calculations of variance estimates. Capacity of today's computers allows everyone to do the calculation of those estimates. Hence, many statistical software packages have been developing procedures that take the complex sample design into account in analyses. Survey Research Center in University of Michigan has done lot of work in following that development. I attended the 61st Summer Institute in Survey Research Techniques there, and this paper presents what I learned about survey procedures in software packages.

1 Introduction

A complex sample survey design produces data with non-independent observations. This must be taken into account in analysis; methods for simple random sample analysis are not valid. In stratified samples the distribution of subpopulation cases is a random variable, and this variability must be considered in subpopulation analysis.

This paper is based on the course material of the Analysis of Complex Sample Survey Data course at the Summer Institute in Survey Research Techniques in the Survey Research Center (SRC). The course was lectured by Steven Heeringa, Brady West and Patricia Berglund. SRC was established at the University of Michigan in 1946 and one of its main functions is to conduct methodological research for the improvement and development of survey procedures. One part of that work has included following the development of survey procedures in statistical software packages and developing their own software for imputation and variance estimation (IVEware). The paper presents survey procedures of some known software packages and introduces features of some not so familiar packages.

2 Complex sample survey data analyzing properties of some statistical software packages

2.1 SAS

SAS versions 9 and higher have four procedures for analysis of complex sample survey data: PROC SURVEYMEANS, PROC SURVEYREG, PROC SURVEYFREQ and PROC

SURVEYLOGISTIC. These four procedures are included in the SAS/STAT package. SAS survey procedures still have limitations in terms of available analyses, for example Poisson regression models cannot be fitted. Hence, count outcome and survival analysis procedures are not available. Also correct calculations for subpopulation regression analyses are still missing in SAS 9.1.3 but should be fixed in SAS 9.2 which is about to be published.

PROC SURVEYMEANS for continuous and binary variables and PROC SURVEYFREQ for categorical variables are for estimating descriptive statistics and their standard errors. For sampling error estimation they use Taylor series linearization method. SAS procedures allow only estimations on the first stage of the sample design.

PROC SURVEYREG is SAS procedure for regression analyses of survey data. It uses Taylor series linearization method for variance estimation. Categorical predictors must be specified with the class keyword and the largest value of the categories defines the default reference category. PROC SURVEYLOGISTIC is procedure for logistic regression and multinomial logistic models analysis of survey data. SAS has no software procedure for Poisson regression analysis of survey data. Weighted Poisson regression can be done with PROC GENMOD, but standard errors will be incorrect.

2.2 STATA

STATA offers very large range of programs for analysis of complex sample survey data. Especially the release 10 includes many new features. For example it supports three different methods of variance estimation; Taylor series linearization, Jackknife repeated replication, and balanced repeated replication methods. It is capable of utilizing information from all stages of the sample design, for example individual, household and municipality levels.

In STATA svyset: -command defines design variables and svydes: -command describes survey design. Those are defined only ones and all the survey procedures that are “svy”-commands take those into account. Svy: mean for continuous and binary variables and svy: prop and svy: tab for categorical variables are commands for estimating descriptive statistics and their standard errors. Svy: regress command performs regression analysis. It allows choice of Taylor series linearization or replication methods for variance estimation. Use of xi: notation before command allows categorical predictors to be identified in the model. Smallest value of the categories will define the reference category.

Logistic regression analysis can be done with svy: logit and svy: logistic commands. STATA has also svylogitgof -command for testing goodness of fit. STATA command for software procedures for Poisson regression analysis of survey data is svy: poisson. Svy: poisson has

also an option for choosing the variance estimation method. This procedure allows modeling a count outcome dependent variable.

2.3 IVEware

IVEware is an imputation and variance estimation software developed by researchers at the Survey Research Center of the Institute for Social Research at University of Michigan. IVEware has both SAS-based and a stand-alone versions and they can be freely downloaded from SRC's website.

Initially IVEware was developed using SAS macros and the structure of SAS procedures is still very much seen. %DESCRIPT, %REGRESS and %SASMOD are its macros for analyzing a complex sample survey data. %DESCRIPT is IVEware's macro for estimating descriptive statistics and their standard errors. Taylor series linearization method is used to estimate the variances and only information about the first stage of the sample design is provided. IVEware output gives many nice features as default; degrees of freedom, design effect, and covariance of denominator for testing the usefulness of Taylor series method.

%REGRESS is macro for regression analysis. It uses jackknife repeated replication for variance estimation. Categorical predictors are specified by categorical statement. The largest value is reference category. By specifying the LINK LOGISTIC option %REGRESS performs logistic regression analysis and LINK LOG performs Poisson regression analysis.

IVEware is well documented: also documentation and examples are found in the SRC's website. See http://www.isr.umich.edu/src/smp/ive.

2.4 SUDAAN and SPSS

Sudan is one of the most known and respected software packages for analysis of correlated data arising from complex sample designs. It is designed at the Research Triangle Institute in the United States. There are SAS-callable and stand-alone versions available. SUDAAN has great variety of survey procedures. All analytic procedures and the new procedure that computes weight adjustments using a model-based, weight calibration methodology offers three variance estimation methods: Taylor series linearization, Jackknife repeated replication, and balanced repeated replication. A new version of SUDAAN has just been released. For more information, visit http://www.rti.org/sudaan/.

SPSS has also released complex samples module for analysis. It is an add-on module that needs to be purchased separately from SPSS base. For more information, visit http://www.spss.com/complex_samples/.

3 Conclusions

All software packages have good features, but also some restrictions that must be taken into account. SUDAAN has probably the most variety of survey procedures available, but is also expensive for individual use. From the "general purpose" statistical software packages STATA offers the most programs for analysis of complex sample survey data and has most precise calculations for variance estimates. SAS survey procedures are included in the SAS/STAT package, but are not so diverse. Even though SAS and STATA offer same statistics sometimes the outputs are not exactly the same. For example, test statistics for categorical data and contingency table analysis, STATA uses a second-order design correction for the Wald F-statistic, while SAS only uses a first-order design correction. Similar difference also appears in their other estimations. IVEware is a free way to SAS user to expand their software package for complex sample survey data analysis. In making the inferences from the analysis, one should always be well aware of the default settings of the package in use.

References

Heeringa, S., West, B. (2008) Analysis of Complex Sample Survey Data. *Course material for the 61st Summer Institute in Survey Research Techniques*. Survey Research Center, University of Michigan.

IVEware documentation, http://www.isr.umich.edu/src/smp/ive.

SAS documentation, http://support.sas.com/documentation/onlinedoc/91pdf/index.html.

SUDAAN documentation, http://www.rti.org/sudaan/.

SPSS documentation, http://www.spss.com/complex_samples/.