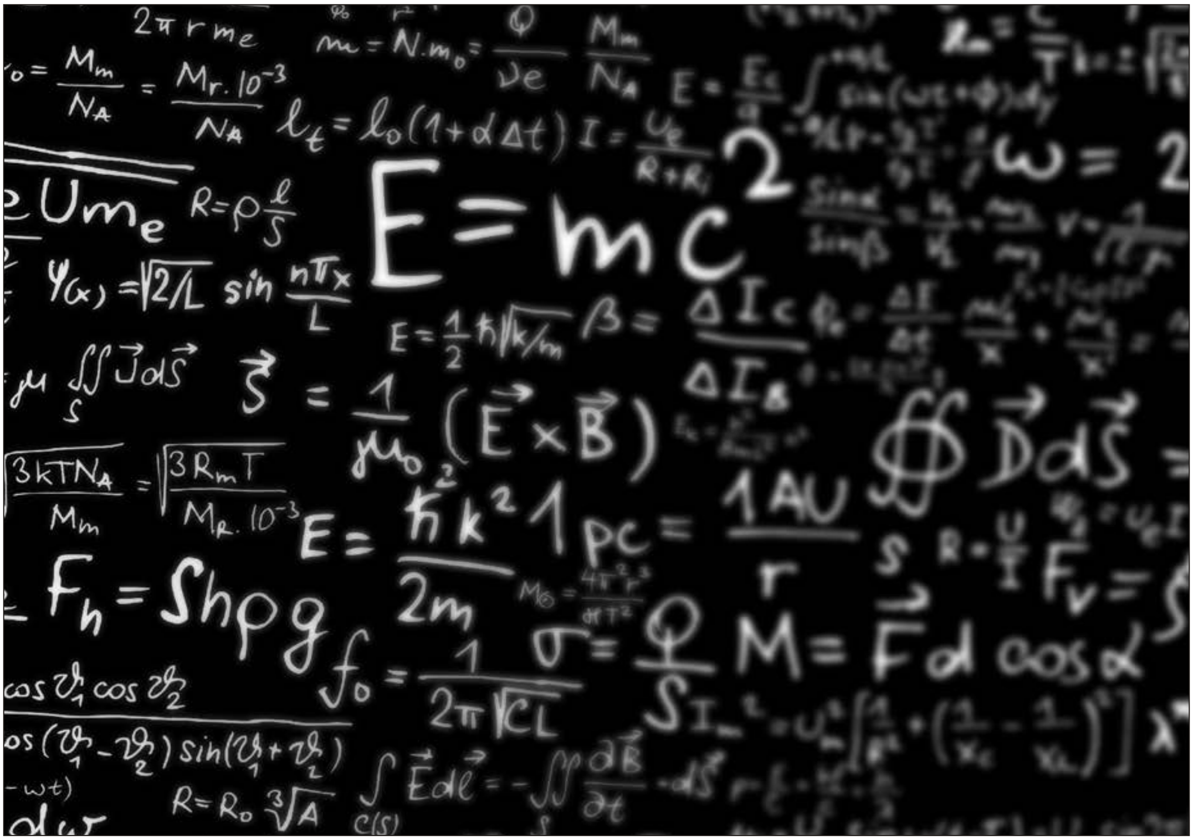

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 3 (1): 33-39, 2010
ISSN 1994-5418

Complex Survey Data Analysis: A Comparison of SAS, SPSS and STATA

[1]G.M. Oyeyemi, [1]A.A. Adewara and [2]R.A. Adeyemi
[1]Department of Statistics, University of Ilorin, P.M.B. 1515 Ilorin, Nigeria
[2]Department of Crop Production,
Federal University of Technology Minna, Minna, Nigeria

Abstract: We compared three statistical packages (SAS, SPSS and STATA) in analyzing complex survey data in the context of multiple regression analysis using concrete examples from two national healthcare database (MEPS and NDHS). The three packages are found to be efficient and flexible in analyzing complex survey data, but SAS in some cases seems to over estimate the variances of the sample statistics. Adjustment for stratification (incorporating stratification) is very important in complex survey analysis, especially if the stratification variable is endogenous.

Key words: Clustering, complex survey, sampling weight, standard error, stratification

INTRODUCTION

A design that is not a simple random sample (where every unit of the target population does not have an equal chance of being selected in the survey) is know as complex survey design. Complex survey sampling is widely used to sample a fraction of large finite population while accounting for its size and characteristics. On the basis of some characteristics of the subject (e.g., age, race, gender etc.) some individuals are over sampled or under sampled. This results in individuals in the population having different probabilities of being selected into the sample (Natarajan *et al.*, 2008).

Demographic and health surveys used for analysis of health sector have complex sample design. In general, sampling is always multistage. Typically, there is simple random sampling at some levels but there might be separate sampling from population subgroups known as strata. Also, there is possibility of group of observations, otherwise known as cluster, which might not be sampled independently and there may be over sampling or under sampling of certain groups. The combination of these different sampling schemes constitutes what is known as a complex design (Oyeyemi *et al.*, 2009).

In this study, we compare analysis of complex survey data in the context of regression analysis using SAS (2002), SPSS (2006) and STATA (2005) statistical software with two different set of complex survey data.

COMPLEX HEALTH SURVEY DESIGNS

The Medical Expenditure Panel Survey (MEPS)

The MEPS is nationally representative survey of the United States civilian non-institutionalized population. It collects medical expenditure data as well as information

Corresponding Author: G.M. Oyeyemi, Department of Statistics, University of Ilorin, P.M.B. 1515 Ilorin, Nigeria

on demographic characteristics, access to health care, health insurance coverage, as well as income and employment data. The MEPS is co-sponsored by the Agency for Healthcare Research and Quality (AHRQ) and the National Centre for Health Statistics (NCHS). For the comparison reported in this study, we used MEPS 2002 (Cohen, 2003). The MEPS is a stratified, multistage probability cluster sample. The data constitute 12583 females who participated in the household component of MEPS. The population was first stratified into 203 geographical regions, known as strata, within each stratum, the geographical region was subdivided into segments, where an area is composed of counties or groups of contiguous counties. The area segments are considered to be clusters or Primary Sampling Units (PSUs) within strata.

Two or three PSUs (area segments) were sampled within each stratum. On the average, a typical PSU contained 60 subjects, with a range of 21-251 subjects. Although, the clusters within strata were sampled without replacement, we can assume that they were sampled with replacement, since, the fraction of clusters that were sample within each stratum is much less than 1% (Natarajan *et al*., 2008). In analyzing these data, one must use the appropriate sampling weights, strata and cluster variables to account for the sampling design (Korn and Grauband, 1999). The outcome of interest is female's total health care expenditure in the year 2002. The covariates of interest are age (age), race (race), smoking status (smoke), level of poverty (pov), health insurance status (insur), self-assessed health status (phealth) and prescription medications (meds).

The Nigeria Demographic and Health Survey (NDHS)

The NDHS provides estimates of national health and family planning statistics. The survey is designed to provide estimates for Nigeria as a whole, for rural and urban areas and for the six geopolitical regions of the country. The survey is always conducted on yearly basis. The 2003 NDHS data will be used in this study. The NDHS is also stratified, multistage probability cluster sample. The six geopolitical zones (North-central, Northeast, Northwest, Southeast, Southwest and South-South) constitute the strata. Within each zone (stratum) there are at least five states with each state subdivided into Local Government Areas (LGAs). The LGAs are composed of wards, these wards are considered to be clusters or primary sampling units PSUs. The PSUs contain clusters of households.

In all, a total of 5,138 subjects were sampled from which various health indicators, socio-economic and other vital statistics were collected. Few variables (health indicators) were considered for the regression analysis for the purpose of comparison in this study. The health used as the dependent variable is the nutritional status of the children and some of selected variables as the socio-economic status indicator (regressors). A child's nutritional status is assessed by comparing the height and weight measurements against an international standard. By this standard, many children in Nigeria are malnourished (NDHS, 2003).

Three categories of nutritional status are identified and they are; stunting (height-for age), wasting (weight-for-height) and underweight (weight-for-age). The socio-economic variables considered are wealth index of parent (wealth), ideal number of children (idlchid), size of the baby at birth (chdsize), current age of the child (currage), sex of the child (sex), parent's index of height for age (reshta), total number of children ever born (totchid), educational level of the mother (educat) and the religion of the parents (religion). The nutritional status used as the dependent variable is stunting.

COMPONENT PARTS OF COMPLEX SURVEY DESIGN

The sampling variance of a survey statistics is affected by the stratification, clustering and weighting of selected cases. Theses three components must be considered in analysis of complex survey data. While stratification may increase the precision of the variance estimated, clustering and weighting normally decrease precision (Dowd and Duggan, 2001).

Stratification

This is a method of using auxiliary variable to increase the precision of the estimate of a population characteristic (Cochran, 1977; Okafor, 2002; Rajj and Chandhok, 1999). Stratification is typically employed in household surveys undertaking in developing countries. Most health surveys use geopolitical zones or regions as the stratification variable. A random sample, of predetermined size, is then selected independently from each of the strata. The sample accounted for by each stratum may or may not correspond to population proportion. There is equal allocation, proportional allocation, optimum allocation among others depending on the design and situation.

In case the sample proportions do not correspond to the population proportions, the overall sample is not representative of the population and the issue of sample weight arises. If the population means differ across the strata, predetermination of strata sample sizes reduces the sampling variance of the estimates of the means (Ferguson and Carey, 1990). Consequently, standard error of estimates of population means and some other statistics should be adjusted. Also, adjustment is not necessary in regression analysis and in wide variety of other multivariate modeling approaches provided stratification is exogenous within the model (Wooldridge, 2001, 2005). Ordinary Least Square (OLS) estimation is found to be consistent and efficient and the usual standard errors are valid in such case. But if stratification is based on endogenous variable then the standard errors should be adjusted (Wooldridge, 2001).

Clustering

In most survey, especially a large and complex survey, there is no sampling frame or list of households or dwelling units from which to select a sample. Therefore, it is not possible or feasible to draw a sample directly from the population. In order to overcome this problem, groups of elements called cluster are formed by pulling together elements which are physically closed to each other (Cochran, 1977; Okafor, 2002). A sample of these cluster units is then selected from the total number of clusters by an appropriate sampling scheme.

Cluster sampling in health and demographic survey has two stages or more sampling processes. In the first stage, groups known as clusters of households are randomly sampled from either the population or strata. Typically, these clusters are villages, hamlets or neighborhood of towns or cities. In the second stage, households are randomly sampled from each of the selected clusters. An important distinction of cluster sampling from stratifying sampling is that strata are selected deterministically, whereas clusters are selected randomly (Owen *et al*., 2007). Another difference is that strata are typically few in number and contain many observations, whereas clusters are large in number but contain relatively few observations.

As a result of this design, observations are expected not to be independent within clusters but may be independent across clusters. There is likely to be more homogeneity within cluster than across clusters, hence, the needs for adjustment or incorporating clustering in the model (estimation of variance components).

Sampling Weight

In most of health and demographic survey, some units may be over sampled or under sampled, therefore, the overall sample is not representative of the population, therefore, different weight are assigned to units sampled. In complex survey design, each unit is selected with an unequal probability of selection and represents a different number of units in the population. The sampling weight assigned to a unit indicates the number of units in the population represented by the respondent.

Weighted estimation equations (Shah *et al*., 1997; Binder, 1983; Pfeffermann, 1993) are the most popular methods for obtaining consistent estimates of the regression coefficient with sample survey data. The contribution to the estimating equation from an individual in the sample survey is weighted by the sample survey weight, which is the inverse of the probability of being selected. For example, if a unit is selected as 100 out of 1000 units, the selection probability is 1\10 and the sampling weight is 10. Most software packages have programs which are not specifically designed for complex sample survey, but which can be used for random-cluster sampling in which individuals within clusters have different weights.

MULTIPLE REGRESSION ANALYSIS USING THE THREE PACKAGES

Multiple regression analysis of the complex survey data starting with MEPS 2002 and then NDHS 2003 data, were done using all the three statistical software packages (SAS, SPSS and STATA) for comparison. For each of the data set, three different models were obtained by incorporating:

- **Model 1:** Stratification, clustering and sample weight in the model
- **Model 2:** Stratification and weight in the model
- **Model 3:** Clustering and weight in the model

Our interest is not on significance of the model coefficients but on their variance estimates in terms of standard error. In all the three different models highlighted above, the three packages gave the same estimates of the regression coefficients but with different estimates of standard errors. The results are shown in Table 1-3 for MEPS 2002 data set and Table 4-6 for NDHS data set for the three models, respectively.

Table 1: Incorporating stratification, clustering and sample weight in the model for MEPS data

		Standard errors		
Variables	Coefficient	SAS	SPSS	STATA
Age	62.835	4.882	4.880	4.880
Smoke	-238.513	170.081	170.033	170.033
Race	498.054	189.279	189.225	189.225
Pov	-445.729	197.505	197.448	197.448
Insur	1926.827	138.237	138.197	138.197
Phealth	-4704.023	329.800	329.706	329.706
Meds	938.175	137.422	137.383	137.383
Constant	2389.619	413.540	413.422	413.422

Table 2: Incorporating stratification and sample weight in the model for MEPS data

		Standard errors		
Variables	Coefficient	SAS	SPSS	STATA
Age	62.835	4.575	4.574	4.574
Smoke	-238.513	185.964	185.911	185.911
Race	498.054	169.675	169.627	169.627
Pov	-445.729	196.781	196.724	196.724
Insur	1926.827	133.411	133.373	133.373
Phealth	-4704.023	328.442	328.348	328.348
Meds	938.175	134.204	134.166	134.166
Constant	2389.619	404.374	404.258	404.258

Table 3: Incorporating clustering and sample weight in the model for MEPS data

		Standard errors		
Variables	Coefficient	SAS	SPSS	STATA
Age	62.835	6.420	6.418	6.418
Smoke	-238.513	186.447	186.394	186.394
Race	498.054	134.644	134.605	134.605
Pov	-445.729	79.724	79.701	79.701
Insur	1926.827	55.425	55.409	55.409
Phealth	-4704.023	385.685	385.575	385.575
Meds	938.175	206.064	206.006	206.006
Constant	2389.619	329.820	329.725	329.725

Table 4: Incorporating stratification, clustering and sample weight in the model for NDHS data

		Standard errors		
Variables	Coefficient	SAS	SPSS	STATA
Wealth	0.169	0.030	0.030	0.030
Idlchid	-0.048	0.011	0.011	0.011
Chdsize	-0.088	0.041	0.041	0.041
Currage	-0.166	0.024	0.024	0.024
Sex	-0.162	0.053	0.053	0.053
Reshta	0.155	0.032	0.032	0.032
Totchid	0.035	0.012	0.012	0.012
Educat	0.194	0.048	0.048	0.048
Religion	-0.566	0.082	0.082	0.082
Constant	-0.377	0.220	0.220	0.220

Table 5: Incorporating stratification and sample weight in the model for NDHS data

		Standard errors		
Variables	Coefficient	SAS	SPSS	STATA
Wealth	0.169	0.022	0.022	0.022
Idlchid	-0.048	0.009	0.009	0.009
Chdsize	-0.088	0.037	0.037	0.037
Currage	-0.166	0.020	0.020	0.020
Sex	-0.162	0.052	0.052	0.052
Reshta	0.155	0.026	0.026	0.026
Totchid	0.035	0.010	0.010	0.010
Educat	0.194	0.039	0.039	0.039
Religion	-0.566	0.060	0.060	0.060
Constant	-0.377	0.162	0.162	0.162

Table 6: Incorporating clustering and sample weight in the model for NDHS data

		Standard errors		
Variables	Coefficient	SAS	SPSS	STATA
Wealth	0.169	0.030	0.030	0.030
Idlchid	-0.048	0.011	0.011	0.011
Chdsize	-0.088	0.041	0.041	0.041
Currage	-0.166	0.024	0.024	0.024
Sex	-0.162	0.053	0.053	0.053
Reshta	0.155	0.032	0.032	0.032
Totchid	0.035	0.012	0.012	0.012
Educat	0.194	0.047	0.047	0.047
Religion	-0.566	0.084	0.084	0.084
Constant	-0.377	0.220	0.220	0.220

For MEPS 2002 data set, the SAS procedure (2002) seems to have over-estimated the standard error than the other two packages (SPSS and STATA). As a matter of fact, both SPSS (2006)and STATA (2005) have the same standard errors in each model.

Interestingly, for all the three packages, the standard errors estimated for the regression coefficients are smaller in the model 2 (when stratification and weight are considered in the model) than the other two models (Table 1-3).

For NDHS 2003 data, all the three packages (SAS, SPSS and STATA) have the same estimated standard errors of the regression coefficients in each model. Also, for all the three packages, except for two or three cases, the estimated standard errors of the regression coefficients are smaller in model 2 (when stratification and weight are considered in the model) than the other two models (Table 4-6). For consistency, all the values are obtained to three places of decimal.

RESULTS AND DISCUSSION

All the three statistical packages are found to be proficient and flexible in analyzing complex survey data. Table 1-3 show the regression coefficients with their standard errors, as obtained by SAS, SPSS and STATA in columns 3, 4 and 5, respectively using MEPS data. While model in Table 1 incorporated stratification, clustering and sample weight, Table 2 incorporated stratification and sample weight. Table 3 has clustering and sample weight in its model. Similarly, Table 4-6 show the results for NDHS data in the like manner as MEPS data.

For all the models (Table 1-6), irrespective of the design or data used, while the SPSS and STATA packages gave exactly the same estimate of standard errors of the regression coefficients, the SAS standard errors are different and in most cases higher than those estimated by the other two packages. Apart from the fact that SAS, in some cases, seems to over-estimated the variance components of the sample statistics, it is more complex to handle for novice and non-statisticians. Similarly, STATA needs commands to execute most of the complex survey analysis which may not be easily grasped by the new users. The SPSS is very easy to handle for the new users or novices than any of other two packages (SAS and STATA), since, one does not need to write commands or programmes to execute any task, all the complex survey analyses can be executed with Graphical User Inter-phase (GUI).

REFERENCES

Binder, D., 1983. On the variance of asymptotically normal estimators from complex surveys. Int. Statist. Rev., 51: 279-292.

Cochran, W.G., 1977. Sampling Techniques. 3rd Edn., Willey Eastern Ltd., New Delhi, India, ISBN: 0852261217.

Cohen, S.B., 2003. Design strategies and innovations in the medical expenditure panel survey. Med. Care, 41: 5-12.

Dowd, A.C. and M.B. Duggan, 2001. Computing Variances from Data with Complex Sampling Designs: A Comparison of Stata and SPSS. North American Stata Users Group, America.

Ferguson, J.A. and P.N. Corey, 1990. Adjusting for clustering in survey design. Ann. Pharmacother., 24: 310-313.

Korn, E.L. and B.I. Grauband, 1999. Analysis of Health Surveys. 1st Edn., Wiley and Sons, New York.

Natarajan, S., S.R. Lipsitz, C.G. Moore and R. Gonin, 2008. Variance estimation in complex survey sampling for generalized linear models. Applied Statistics, 57: 75-87.

NDHS., 2003. National Population Commission [Nigeria] and ORC Macro. NDHS., USA.

Okafor, C.F., 2002. Sample Survey Theory with Applications. 1st Edn., Afro-Orbis Publications Ltd., Nsukka, Nigeria.

Owen, O.D., V.D. Eddy, W. Adam and L. Magnus, 2007. Analyzing Health Equity Using Household Survey Data: A Guide to Techniques and their Implementation. World Bank Publications, USA.

Oyeyemi, G.M., A.A. Adewara, B.A. Ibraheem and S.O. Ige, 2009. Multivariate Regression in Complex Survey Design. ICASTR, Indian.

Pfeffermann, D., 1993. the role of sampling weights when modeling survey data. Int. Statist. Rev., 61: 317-337.

Rajj, D. and P. Chandhok, 1999. Sample Survey Theory. 1st Edn., Nawsa Publication House, London.

SAS., 2002. SAS Systems for Windows 9.0. SAS Institute Inc., Cary, NC, USA.

Shah, B.V., B.G. Barnwell and G.S. Bieler, 1997. Sudaan Users Manual, Release 7.5. Research Triangle Institute, USA.

SPSS., 2006. SPSS for Windows 15.0. SPSS Inc., USA.

Stata Corporation, 2005. STATA Release 9 Survey Data Reference Manual. Stata Corporation, USA.

Wooldridge, J.M., 2001. On the Robustness of Fixed Effects and Related Estimators in Correlated Random Coefficients Panel Data Models. Institute for Fiscal Studies Series, London.

Wooldridge, J.M., 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved effects. J. Applied Econometrics, 20: 39-54.