Using Statistical Software Packages to Produce Estimates from MEPS Data File

Introduction

The Household Component of the Medical Expenditure Panel Survey (MEPS-HC) is designed to produce national and regional estimates of the health care use, expenditures, sources of payment, and insurance coverage of the U.S. civilian non-institutionalized population. The sample design of the survey includes stratification, clustering, multiple stages of selection, and disproportionate sampling. Furthermore, the MEPS sampling weights reflect adjustments for survey nonresponse and adjustments to population control totals from the Current Population Survey. These survey design and estimation complexities require special consideration when analyzing MEPS data (i.e., it is not appropriate to assume simple random sampling).

To obtain accurate estimates from MEPS survey data, for either descriptive statistics or more sophisticated analyses based on multivariate models, the MEPS survey design complexities need to be taken into account by applying MEPS survey weights to produce estimates and using an appropriate technique to derive standard errors associated with the weighted estimates. Several methods for estimating standard errors for estimates from complex surveys have been developed, including the Taylor-series linearization method, balanced repeated replication, and the jack-knife method.

The MEPS public use files include variables to obtain weighted estimates and to implement a Taylor-series approach to estimate standard errors for weighted survey estimates. These variables, which jointly reflect the MEPS survey design, include the survey weight, sampling strata, and primary sampling unit (PSU). The documentation and codebook for MEPS public use files contain these survey design variables. For example, the documentation for file HC-147 (2011 full-year consolidated data file) includes the person weight (PERWT11F), stratum (VARSTR), and PSU (VARPSU) variables.

Statistical software packages that are commonly used to estimate standard errors from complex multistage designs using the Taylor-series linearization method include SAS® (version 8.2 or higher), SUDAAN®, SPSS® (version 12.0 or higher), Stata® and the R survey package. Examples of basic programming code from these packages to produce selected estimates and the corresponding standard errors are provided in this document. The software packages vary with respect to the specific types of estimates and models that can be produced accounting for the complex survey design and the treatment of missing data. For complete information on the capabilities of each package, analysts need to refer to the appropriate software user documentation manuals. The Web sites for these packages are contained in the table below.

Software Package	**Website**
SAS	sas.com
SUDAAN	rti.org
SPSS	spss.com
Stata	stata.com
R survey	r-survey.r-forge.r-project.org/survey/

Examples

The tables below provide examples of basic programming code for the software packages (SAS, SUDAAN, and SPSS in the first table; Stata and R in the second table) to generate estimates from MEPS person-level files for 1) the total population, 2) population subgroups and 3) differences between population subgroups. Following is a brief description of the examples illustrated in the tables.

1. **Total Population**

   Using the 2011 MEPS full-year consolidated file (PUF HC-147) as the analytic file, the basic programming code provided for each software package in the table below will produce correct estimates of the overall mean total expenditures in 2011 ($4,277.13) and the corresponding standard error ($111.14).

2. **Population Subgroup**

   Analyses are often limited to a subgroup of the population. However, creating a special analysis file that contains only observations for the subgroup of interest may yield incorrect standard errors or an error message (e.g., "stratum with only one PSU detected" in Stata) because all of the observations corresponding to a stage of the MEPS sample design may be deleted. Therefore, it is advisable to preserve the entire survey design structure for the program by reading in the entire person-level file. Each software package provides a capability to limit the analysis to a subgroup of the population without sub-setting the analysis file.

   Using the 2011 MEPS full year consolidated file (PUF HC-147) as the analytic file, the basic programming code provided for each software package in the table below will produce accurate estimates of the average total expenditures in 2002 for children younger than 18 years of age ($1,586.82) and the corresponding standard error ($84.95).

3. **Difference Between Subgroups**

   Analysts often need to test whether differences between subdomains are statistically significant. A specific example comparing males to females in MEPS 2011 is provided to illustrate the appropriate syntax for the five software packages.

   Using the 2011 MEPS full year consolidated file (PUF HC-147) as the analytic file, the mean total expenditures for males was $3,967.09, the mean total expenditures for females was $4,573.49, with a difference of -$606.40. The difference has a standard error of $211.42 with a t-value of -2.87 and a p-value of 0.005.

Using Statistical Software Packages to Produce Estimates from MEPS Data File

Using SAS, SUDAAN, and SPSS to Compute Standard Errors for MEPS Estimates (Examples based on Public Use File HC-147)

		SAS	SUDAAN	SPSS
Full population	**Code**	PROC SURVEYMEANS DATA=FY; STRATUM VARSTR; CLUSTER VARPSU; WEIGHT PERWT11F; VAR TOTEXP11; RUN;	PROC DESCRIPT FILETYPE=SAS DESIGN=WR DATA=FY; NEST VARSTR VARPSU /MISSUNIT; WEIGHT PERWT11F; VAR TOTEXP11; PRINT MEAN SEMEAN / MEANFMT=F12.4 SEMEANFMT=F12.4; RUN;	CSPLAN ANALYSIS /PLAN FILE ="C:\mepsdsgn.csplan" /PLANVARS ANALYSISWEIGHT=PERWT11F /DESIGN STRATA =VARSTR CLUSTER =VARPSU /ESTIMATOR TYPE =WR. CSDESCRIPTIVES /PLAN FILE ="C:\mepsdsgn.csplan" /SUMMARY VARIABLES =totexp11 /MEAN /STATISTICS SE.
	Mean	$ 4,277.13	$ 4,277.13	$ 4,277.13
	SE	$ 111.14	$ 111.14	$ 111.14
Subpopulation	**Code**	PROC SURVEYMEANS DATA=FY; STRATUM VARSTR; CLUSTER VARPSU; WEIGHT PERWT11F; VAR TOTEXP11; DOMAIN CHILD; RUN;	PROC DESCRIPT FILETYPE=SAS DESIGN=WR DATA=FY; NEST VARSTR VARPSU /MISSUNIT; WEIGHT PERWT11F; VAR TOTEXP11; PRINT MEAN SEMEAN / MEANFMT=F12.4 SEMEANFMT=F12.4; SUBPOPN CHILD=1; RUN;	CSPLAN ANALYSIS /PLAN FILE ="C:\mepsdsgn.csplan" /PLANVARS ANALYSISWEIGHT=PERWT11F /DESIGN STRATA =VARSTR CLUSTER =VARPSU /ESTIMATOR TYPE =WR. CSDESCRIPTIVES /PLAN FILE ="C:\mepsdsgn.csplan" /SUMMARY VARIABLES =totexp11 /SUBPOP TABLE = CHILD /MEAN /STATISTICS SE.
	Mean	$ 1,586.82	$ 1,586.82	$ 1,586.82
	SE	$ 84.95	$ 84.95	$ 84.95
Comparison [a] between male and female	**Code**	PROC SURVEYREG DATA=FY; STRATUM VARSTR; CLUSTER VARPSU; WEIGHT PERWT11F; CLASS SEX; MODEL TOTEXP11=SEX/NOINT SOLUTION VADJUST=NONE; LSMEANS SEX/DIFF; CONTRAST "Compare male vs. female" SEX 1 -1; RUN;	PROC DESCRIPT FILETYPE=SAS DESIGN=WR DATA=FY; NEST VARSTR VARPSU /MISSUNIT; WEIGHT PERWT11F; VAR TOTEXP11; CLASS SEX; PAIRWISE SEX; CONTRAST SEX=(1 -1); DIFFVAR SEX=(1 2); PRINT MEAN SEMEAN T_MEAN P_MEAN /MEANFMT=F12.4 SEMEANFMT=F12.4 T_MEANFMT=F12.4 P_MEANFMT=F12.4; RUN;	CSGLM TOTEXP11 BY SEX /PLAN FILE ="C:\mepsdsgn.csplan" /MODEL SEX /INTERCEPT INCLUDE =NO /TEST TYPE=F /EMMEANS TABLES =SEX COMPARE=SEX CONTRAST=SIMPLE /CRITERIA CILEVEL =95.
	Mean for male	$ 3,967.09	$ 3,967.09	$ 3,967.09
	SE for male	$ 183.18	$ 183.18	$ 183.18
	Mean for female	$ 4,573.49	$ 4,573.49	$ 4,573.49
	SE for female	$ 117.88	$ 117.88	$ 117.88
	Difference	-$ 606.40	-$ 606.40	-$ 606.40
	SE of difference	$ 211.42	$ 211.42	$ 211.42
	t value	-2.870	-2.868	[b]
	p	0.005	0.005	[b]
	Wald F value	8.230	[b]	8.227
	p	0.005	[b]	0.005

[a] Notes on comparison:

- SAS will not produce a result for a single subdomain only
- SUDAAN, in addition to PAIRWISE, DIFFVAR and CONTRAST can also be used to generate two-way comparisons.

[b] Statistic not produced by this software package.

Using STATA and R to Compute Standard Errors for MEPS Estimates (Examples based on Public Use File HC-147)

		STATA	R
Full population	**Code**	svyset [pweight=perwt11f], strata(varstr) psu(varpsu) svy: mean totexp11	library(survey) options(digits=10) mepsdsgn <- svydesign(id=~VARPSU, strata=~VARSTR, weights=~PERWT11F, data=FY, nest=TRUE) svymean(~TOTEXP11, mepsdsgn)
	Mean	$ 4,277.13	$ 4,277.13
	SE	$ 111.14	$ 111.14
Subpopulation	**Code**	svy: mean totexp11, subpop(child)	svymean(~TOTEXP11, subset(mepsdsgn, CHILD==1))
	Mean	$ 1,586.82	$ 1,586.82
	SE	$ 84.95	$ 84.95

Comparison between male and female	**Code**	svy: mean totexp11, over(sex)
		lincom [totexp11]1-[totexp11]2
			svyby(~TOTEXP11, ~SEX, mepsdsgn, svymean)
			svyttest(TOTEXP11~SEX, mepsdsgn)
			
			# alternative way
			summary(svyglm(TOTEXP11~ factor(SEX), mepsdsgn))

	Mean for male	$ 3,967.09	$ 3,967.09
	SE for male	$ 183.18	$ 183.18
	Mean for female	$ 4,573.49	$ 4,573.49
	SE for female	$ 117.88	$ 117.88
	Difference	-$ 606.40	$ 606.40
	SE of difference	$ 211.42	$ 211.42
	t value	-2.870	2.868
	p	0.005	0.005
	Wald F value	[c]	[c]
	p	[c]	[c]

[c] Statistic not produced by this software package.