PSID A national study of socioeconomics and health over lifetimes and across generations

Technical Series Paper #11-05

Sampling Error Estimation in Design-Based Analysis of the PSID Data

Steven G. Heeringa, Patricia A. Berglund, Azam Khan
Survey Research Center, Institute for Social Research
University of Michigan

August, 2011

This project was supported by funding from the National Science Foundation (SES 0518943).

PSID Technical Report

Sampling Error Estimation in Design-based Analysis of the PSID Data.

August 16, 2011

Steven G. Heeringa, Patricia A. Berglund, Azam Khan

University of Michigan, Ann Arbor, MI

This document describes how sampling error estimation and construction of confidence intervals for survey estimates of descriptive statistics such as means, proportions, ratios, and coefficients for linear and logistic regression models can be undertaken using the Panel Study of Income Dynamics (PSID) data. This technical report is organized in four sections. Section I provides an overview of the PSID sample and its complex design. Section II provides an overview of sampling error computation methods and software programs currently in use, focusing on the most commonly used methods (i.e. the Taylor Series Linearization Method and resampling methods such as Balanced Repeated Replication (BRR) and Jackknife Repeated Replication (JRR)) and software programs (i.e. SAS Version 8+, Stata Release 9+, SPSS Version 14+, SUDAAN Version 9+, WesVar Version 4+, and IVEWare). Section III introduces the sampling error computation model for the PSID core and immigrant samples. The report concludes in Section IV with syntax that should be used for design based variance estimation when using SAS, Stata, SPSS, and SUDAAN software programs.

I. The PSID Sample

The PSID is the longest running, nationally representative household panel survey in the world. PSID data have been collected annually from 1968 to 1997 and biennially since 1997. The PSID panel is based on the dynamic longitudinal follow up of families and their descendants originally identified in a combination of three probability samples of U.S. households: the Survey Research Center 1960 National Sample (SRC), a subsample of families interviewed in 1967 by the Bureau of the Census for the Office of Economic Opportunity (SEO) (McGonagle and Schoeni, 2006) and the 1997 PSID Immigrant Supplement (Heeringa and Connor, 1998). Sample persons and their descendents identified in the baseline SRC and SEO samples (termed the PSID "Core" in many publications) have been interviewed since 1968. In 1997 and 1999, the baseline sample of the post-1968 immigrants was added and these new immigrant sample persons have been followed continuously since the late 1990s. More detailed information on the PSID 1968 and 1997/1999 immigrant samples is available from the PSID website, (http://psidonline.isr.umich.edu/). The PSID sample design is similar in its basic structure to the multi-stage designs used for major survey programs. The survey literature refers to the these samples as complex designs, a loosely-used term meant to denote the fact that the sample

incorporates special design features such as stratification, clustering and differential selection probabilities (i.e., weighting) that analysts must consider in computing sampling errors for sample estimates of descriptive statistics and model parameters. Standard programs in statistical analysis software packages assume simple random sampling (SRS) or independence of observations in computing standard errors for sample estimates. In general, the SRS assumption results in underestimation of variances of survey estimates of descriptive statistics and model parameters. Confidence intervals based on computed variances that assume independence of observations will be biased (generally too narrow) and design-based inferences will be affected accordingly. Likewise, test statistics (t, X^2, F) computed in complex survey data analysis using standard programs will tend to be biased upward and overstate the significance of tests of effects.

II. Sampling Error Computation Methods and Programs

Over the past 50 years, advances in survey sampling theory have guided the development of a number of methods for correctly estimating variances from complex sample data sets. A number of sampling error programs that implement these complex sample variance estimation methods are available to PSID data analysts. The two most common approaches (Rust, 1985; Wolter, 1985) to the estimation of sampling error for complex sample data are through the use of a Taylor Series linearization of the estimator (and corresponding approximation to its variance) or through the use of resampling variance estimation procedures such as Balanced Repeated Replication (BRR) or Jackknife Repeated Replication (JRR).

Taylor Series Linearization Method

When survey data are collected using a complex sample design with unequal size clusters, most statistics of interest will not be simple linear functions of the observed data. The linearization approach applies Taylor's method to derive an approximate form of the estimator that is linear in statistics for which variances and covariances can be directly and easily estimated. Stata Releases 9 to 12, SAS V8-V9, SUDAAN Versions 9 and 10 and the most recent releases of SPSS are commercially available statistical software packages that include procedures that apply the Taylor series method to estimation and inference for complex sample data.

Stata (StataCorp, 2007) is a more recent commercial entry to the available software for analysis of complex sample survey data and has a growing body of research users. Stata includes special versions of its standard analysis routines that are designed for the analysis of complex sample survey data. Special survey analysis programs are available for descriptive estimation of means and ratios (svy: mean , ratios (svy: ratio), proportions (svy: proportion) and population totals (svy: total). Stata programs for multivariate analysis of survey data include linear regression (svy: regress), logistic regression (svy: logit and svy: logistic) and probit regression (svy: probit). In addition, there are numerous other svy: commands. Stata program offerings for survey data analysts are constantly being expanded. Information on the Stata analysis software system can be found on the Web at: http://www.stata.com.

Survey analysis procedures were introduced in SAS Version 8 (SAS, 2004; www.sas.com) and also use the Taylor Series method and the optional repeated replication (JRR/BRR) methods to estimate complex sample variances. The current version of SAS 9.3 offers procedures for means (PROC SURVEYMEANS), proportions and cross-tabular analysis (PROC SURVEYFREQ), linear regression (PROC SURVEYREG), logistic regression (PROC SURVEYLOGISTIC) and PROC SURVEYPHREG (new in version 9.22) for Cox Proportional Hazards models .

SUDAAN (RTI, 2004) is a commercially available software system developed and marketed by the Research Triangle Institute of Research Triangle Park, North Carolina (USA). SUDAAN was developed as a stand-alone software system with capabilities for the more important methods for descriptive and multivariate analysis of survey data, including: estimation and inference for means, proportions and rates (PROC DESCRIPT and PROC RATIO); contingency table analysis (PROC CROSSTAB); linear regression (PROC REGRESS); logistic regression (PROC LOGIST/RLOGIST); log-linear models (PROC MULTILOG; and survival analysis (PROC SURVIVAL). SUDAAN V9.0 and earlier versions were designed to read directly from ASCII and SAS system data sets. The latest versions of SUDAAN permit procedures to be called directly from the SAS system. Information on SUDAAN is available at the following web site address: www.rti.org.

SPSS Version 19.0 (http://www-01.ibm.com/software/analytics/spss) users can obtain the SPSS Complex Samples module which supports Taylor Series Linearization estimation of sampling errors for descriptive statistics (CSDESCRIPTIVES), frequencies (CSFREQUENCIES), cross-tabulated data (CSTABULATE), ratios (CSRATIO), general linear models (CSGLM), ordinal logistic regression (CSORDINAL), Cox regression and Kaplan-Meier curves (CSCOXREG), and logistic regression (CSLOGISTIC).

Resampling Methods

BRR, JRR and the bootstrap comprise a second class of nonparametric methods for conducting estimation and inference from complex sample data. As suggested by the generic label for this class of methods, BRR, JRR and the bootstrap utilize replicated subsampling of the sample database to develop sampling variance estimates for linear and nonlinear statistics. WesVar PC (Westat, Inc., 2000) is a software system for personal computers that employs replicated variance estimation methods to conduct the more common types of statistical analysis of complex sample survey data. WesVar PC was developed by Westat, Inc. and is distributed along with documentation to researchers at Westat's Web site: http://www.westat.com/wesvarpc/ . WesVar PC includes a Windows-based application generator that enables the analyst to select the form of data input (SAS data file, SPSS for Windows data base, ASCII data set) and the computation method (BRR or JRR methods). Analysis programs contained in WesVar PC provide the capability for basic descriptive (means, proportions, totals, cross tabulations) and regression (linear, logistic) analysis of complex sample survey data. WesVar also provides an excellent facility for estimating quantiles of continuous variables (e.g. 95%-tile of a cognitive test score) from survey data. WesVar Complex Samples 5.1 is the latest version of WesVar.

Stata V9 and later releases have introduced the option to use JRR or BRR calculation methods as an alternative to the Taylor Series method for all of its svy command options. Beginning with SAS v9.1, JRR or BRR methods are available in each of the SAS SURVEY procedures. SUDAAN V9.0/V10.0 also allows the analyst to select the JRR method for computing sampling variances of survey estimates. SPSS does not offer repeated replication ability at this point.

IVEware (Imputation and Variance Estimation Software) is another software option for JRR estimation of sampling errors for survey statistics. IVEware has been developed by the Survey Methodology Program of the Survey Research Center and is available free of charge to users at: http://www.isr.umich.edu/src/smp/ive/ . IVEware is based on SAS Macros and requires SAS Version 6.12 or higher. The system includes programs for multiple imputation of item missing data as well as programs for variance estimation in descriptive (means and proportions) and multivariate (linear regression, logistic regression, Poisson regression, and survival analysis) analysis of complex sample survey data (Raghunathan, et al., 2001).

Each of the software tools described in the previous section include an expanded set of user friendly, well documented analysis procedures. Difficulties with sample design specification, data preparation, and data input in the earlier generations of survey analysis software created a barrier to use by analysts who were not survey design specialists. The new software packages enable the user to input data and output results in a variety of common formats, and the latest versions accommodate direct input of data files from the major analysis software systems.

For more information on survey data analysis and software tools see "Applied Survey Data Analysis" by Heeringa, West, and Berglund (2010) and the companion website: http://www.isr.umich.edu/src/smp/asda/.

III. Sampling Error Computation Models

Regardless of whether the linearization method or a resampling approach is used, estimation of variances for complex sample survey estimates requires the specification of a *sampling error computation model.* PSID data analysts who are interested in performing sampling error computations should be aware that the estimation programs identified in the preceding section assume a specific sampling error computation model and will require special sampling error codes. Individual records in the analysis data set must be assigned sampling error codes that identify to the programs the complex structure of the sample (stratification, clustering) and are compatible with the computation algorithms of the various programs. To facilitate the computation of sampling error for statistics based on PSID data, design-specific sampling error codes will be routinely included in all public-use versions of the data set. Although minor recoding may be required to conform to the input requirements of the individual programs, the sampling error codes that are provided should enable analysts to conduct either Taylor Series or repeated replication estimation of sampling errors for survey statistics. In programs that use the Taylor Series Linearization method, the sampling error codes (stratum and cluster) will typically be input as keyword statements (SAS V9.1 or higher, SUDAAN V9.0 or higher) or as global settings (Stata V9 or higher, SPSS "plan file") and will be used directly in the computational algorithms. Programs that permit BRR or JRR computations will require the user supplied

sampling error codes to construct “replicate weights” that are required for these approaches to variance estimation or the data producer supplied "replicate weights".

Two sampling error code variables are defined for each case based on the sample design stratum and primary stage unit (PSU) cluster in which the sample respondent resided: Sampling Error Stratum Code (SESTRATUM) and Sampling Error Cluster Code (SECLUSTER). The sampling error SESTRATUM variable for PSID contains a unique code for each of 87 sampling error strata formed by either combining matched pairs of sampling design strata or creating separate sampling error strata for each self-representing primary stage stratum in the parent PSID design.

In variance estimation for complex sample designs, the sampling error clusters represent the “ultimate clusters” (Kalton, 1977) of the sample selection process. The SECLUSTER code reflects the geographic clustering of sample observations based on the PSUs to which they are assigned. Sampling variances for survey estimates will therefore be estimated under the assumption that two PSU sampling error clusters were selected from each stratum.

Table 1 summarizes the distribution of PSID respondent cases to the assigned sampling error calculation strata (SESTRATUM) and clusters (SECLUSTER).

IV. Syntax for PSID Design-based Variance Estimation Using Stata, SAS, SUDAAN and SPSS

The following two sections provide a short overview of the general syntax and command file structure for computing sampling errors using Stata, SAS, SUDAAN, and SPSS programs that have been designed for the analysis of complex sample survey data. Analysts are referred to the user guides and the on-line help facilities of these four software systems for documentation of the individual programs.

Stata Command Syntax

As described above, PSID data analysts who are familiar with the Stata software system can utilize “svy” commands for the analysis of complex sample survey data. Stata 9 and higher syntax for some of the more commonly used analysis programs is illustrated below:

svyset SECLUSTER [pweight= (*PSID WEIGHT OF CHOICE*], strata(SESTRATUM)

This statement defines the sample design variables for the duration of the analysis session. Any "svy" commands issued after this statement will automatically incorporate these design specifications.

To conduct analyses, the following Stata commands and syntax are used (please refer to Stata V9 or higher Reference Manual for specific command syntax and output options):

svy, vce(linearized): mean *vars*
[estimates, standard errors, design effects for means]

svy, vce(linearized): tab *v1 v2*

[estimates, standard errors for proportions of single variable categories, or crosstabulations of two variables with tests of independence]

svy, vce(linearized): regress dep x1 ...
[simple linear regression model for a continuous dependent variable]

svy, vce(linearized): logit dep x1… (for coefficients or log odds in output)
svy, vce(linearized): logistic dep x1... (for Odds Ratios in output)
[simple logistic regression model for a binary dependent variable]

To estimate simple descriptive statistics or regression models for subpopulations of the survey population in STATA, the following optional syntax is used (illustrated for svy, tab):

svy, vce(linearized): tab *v1 v2*, subpopulation*(var)*
[where *var* is a binary variable that equals "1" for the subpopulations for which separate estimates are desired (e.g. males) and "0" for all other cases.]

There are numerous other svy: commands in Stata. See the Stata documentation or the ASDA website for examples.

SAS Command Syntax

SAS Version 9.2 and higher includes five procedures for the analysis of complex sample survey data: PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC and PROC SURVEYPHREG. The general syntax for specifying the PSID design structure in the SAS system is as follows:

```
PROC SURVEYMEANS;
STRATA   SESTRATUM;
CLUSTER SECLUSTER;
WEIGHT (PSID WEIGHT OF CHOICE);
additional program specific statements here;
RUN;
```

Users are referred to the current SAS/STAT® *User's Guide* or SAS On-line Help for documentation on program specific statements, keywords and options.

SUDAAN Command Syntax

SUDAAN (all versions) includes numerous procedures for the analysis of complex sample survey data: PROC DESCRIPT, PROC CROSSTAB, PROC RATIO, PROC REGRESS, PROC LOGIST (PROC RLOGIST for SAS callable SUDAAN), PROC MULTILOG, PROC LOGLINK, PROC SURVIVAL and SURVEYPHREG.

The general syntax for specifying the PSID design structure in the SUDAAN system is as follows:

```
PROC DESCRIPT design=wr ;
NEST SESTRATUM SECLUSTER;
WEIGHT (PSID WEIGHT  OF CHOICE) ;
additional program specific statements here;
RUN;
```

Users are referred to the current SUDAAN Language and Examples Manuals for additional help.

SPSS Command Syntax

SPSS (versions 14+) includes numerous procedures for the analysis of complex sample survey data as part of the Complex Samples Module: CSDESCRIPTIVES, CSTABULATE, CSFREQUENCIES,CSRATIOS, CSGLM, CSLOGISTIC, CSORDINAL, and CSCOXREG.

The general syntax for specifying the PSID design structure in the SPSS system is done within the CSPLAN file, which is set up prior to analysis.

```
CSPLAN ANALYSIS
  /PLAN FILE='(PATH OF LOCATION AND NAME OF PLAN FILE)'
  /PLANVARS ANALYSISWEIGHT=( PSID WEIGHT OF CHOICE)
  /SRSESTIMATOR TYPE=WR
  /PRINT PLAN
  /DESIGN STRATA=SESTRATUM CLUSTER=SECLUSTER
  /ESTIMATOR TYPE=WR.
```

Once the plan file is set with the correct complex sample variables, the analysis the syntax would be similar to the following:

```
 CSDESCRIPTIVES
/PLAN FILE='(INSERT PATH/NAME OF PLAN FILE HERE)'
/SUMMARY VARIABLES=AGE
additional program specific statements here
```

Users are referred to the current SPSS Manuals and online tutorials for additional help.

V. References

Heeringa, S.G. and Connor, J.H. (1997). "Technical documentation for the 1997 PSID Sample". Panel Study of Income Dynamics Technical Report. Survey Research Center, University of Michigan, Ann Arbor.

Heeringa, S.G. and Connor J.H. (1998). "Technical documentation for the 1997 PSID Immigrant Supplement". Panel Study of Income Dynamics Technical Report. Survey Research Center, University of Michigan, Ann Arbor.

Heeringa, S.G., Berglund, P.A., Khan, A. (2011). "Panel Study of Income Dynamics. Construction and Evaluation of the Longitudinal Individual and Family Weights" . Panel Study of Income Dynamics Technical Report. Survey Research Center, University of Michigan, Ann Arbor.

Heeringa, S.G., West, B.T., Berglund, P.A. (2010). *Applied Survey Data Analysis*. Chapman and Hall (CRC).

Hill, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide.* Newbury Park, CA: Sage Publications.

Kalton, G. (1977), "Practical Methods for estimating survey sampling errors", *Bulletin of the International Statistical Institute,* Vol 47, 3, pp. 495-514.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

McGonagle, K. and Schoeni, R. (2006). "The Panel Study of Income Dynamics: Overview and Summary of Scientific Contributions After Nearly 40 Years." Panel Study of Income Dynamics Technical Paper Series. Available at: http://psidonline.isr.umich.edu/Publications/Papers/tsp/2006-01_PSID_Overview_and_summary_40_years.pdf

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P. (2001). "A Multivariate Technique for Multiple Imputation Using a Sequence of Regression Models.", *Survey Methodology,* Vol. 27, No.1.

Research Triangle Institute. (2004). SUDAAN 9.0 User's Manual: Software for Statistical Analysis of Correlated Data. Research Triangle Park, NC: Research Triangle Institute.

Rust, K. (1985). "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics*, Vol. 1, No. 4.

SAS Institute, Inc. (2003). SAS/STAT® User's Guide, Version 9, Cary, NC: SAS Institute, Inc.

STATA Corp. (2005). STATA Release 10.0-Survey Data. College Station, TX: STATA Corporation.

Westat, Inc. (2000). WesVar 4.0 User's Guide. Rockville, MD: Westat, Inc.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Table 1. Distribution of PSID Respondents by Sampling Error Stratum and Cluster

SESTRATUM	SECLUSTER	
	1	2
1	1389	2073
2	2326	1730
3	744	939
4	487	799
5	1382	1097
6	1127	1131
7	1544	1486
8	1009	943
9	1209	1254
10	416	483
11	589	413
12	551	979
13	588	811
14	522	594
15	874	1191
16	295	813
17	604	658
18	434	530
19	890	721
20	927	1164
21	875	662
22	1638	1357
23	1014	1013
24	840	920
25	662	1177
26	496	475
27	532	448
28	512	604
29	524	720
30	993	1032
31	803	1177
32	532	928
33	113	118
34	184	154
35	336	442
36	273	313
37	252	252
38	331	295

SESTRATUM	SECLUSTER	
	1	2
39	132	87
40	73	123
41	101	91
42	94	93
43	295	178
44	182	130
45	230	145
46	156	112
47	239	138
48	184	163
49	215	121
50	241	156
51	254	305
52	268	316
53	296	167
54	891	430
55	260	233
56	245	200
57	4	15
58	69	63
59	54	44
60	24	43
61	43	28
62	42	40
63	42	28
64	45	54
65	42	52
66	28	69
67	7	7
68	36	31
69	34	29
70	54	46
71	23	21
72	42	31
73	35	42
74	25	8
75	75	121
76	31	36
77	8	11

SESTRATUM	SECLUSTER	
	1	2
78	98	46
79	42	41
80	95	236
81	16	28
82	82	63
83	75	89
84	54	44
85	68	56
86	55	45
87	108	105
Total	**34629**	**36656**