THE UNIVERSITY OF
MELBOURNE

HILDA PROJECT TECHNICAL PAPER SERIES
No. 2/08, February 2008

HILDA Standard Errors: A Users Guide

Clinton Hayes

The HILDA Project was initiated, and is funded, by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs

MELBOURNE INSTITUTE
of Applied Economic and Social Research

ACER

Australian Government
Australian Institute of Family Studies

Contents

HILDA Survey Design **1**
Comparison to a Simple Random Sample *1*
Standard Errors for Complex Surveys **2**
Taylor Series Linearisation *2*
Delete-a-Group Jackknife *3*
Data **5**
Software **6**
STATA *6*
Examples *6*
SPSS *7*
Examples *8*
SAS *8*
Examples *9*
References **12**

HILDA Survey Design

The Household, Income and Labour Dynamics in Australia (HILDA) Survey is designed to be representative of Australia. We could have implemented simple random sampling (SRS) to achieve this aim, but it is a very expensive way of collecting data. Instead, a complex survey design was used to select households in wave 1. In addition to cost benefits, a complex survey design allows for improved estimates at an area level by ensuring that a suitable number of units are selected within each. The HILDA selection was performed in the following order:

- Australia was stratified via major statistical region and Census Collection Districts (CD’s) were systematically selected (via serpentine ordering) within each strata, with probability proportional to the number of dwellings in each CD;
- Dwellings were selected systematically from within CD’s;
- Households were randomly selected within dwellings; and
- All individuals were selected in each household.

The sample design aimed to give close to equal weights to all individuals, but in reality, the weights differ mainly due to non-response, but also due to inaccurate information on the number of dwellings in each CD at the time they were selected.

Comparison to a Simple Random Sample

A common error when dealing with a complex survey is to assume that SRS formulae can be used to estimate variances. This is incorrect and any estimate of variance must take into account the sampling design.

The geographical nature of the survey design results in selected CD’s (clusters) being well disbursed throughout each stratum while households within each cluster tend to be very similar. The clustering of households, and the clustering of people within households, may cause SRS formulae to produce a biased estimate of the true error. The direction of the bias can be uncertain as it is dependent on how the variable being estimated is related to geographical area. Estimates of the variance need to be adjusted for the effect of clustering and stratification to produce reliable results.

Standard Errors for Complex Surveys

There are many methods available to handle standard error calculation for complex surveys. The HILDA User Manual[1] suggests the Taylor Series linearization method and the Jackknife delete-a-group replicate method, so the focus on this paper will be to briefly present these two methods.

Taylor Series Linearisation

The Taylor Series expansion is typically used to obtain an approximation to a function that is hard to calculate. For instance, the Taylor expansion of e^x is produced by taking the 1st and higher order derivatives of e^x with respect to x; evaluating the derivatives at some value, usually zero; and building up a series of terms based of the derivatives. The expansion for e^x is

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \tag{1}$$

A statistic of interest from a complex survey is non-linear and the formula for the variance of a non-linear statistic is very difficult to solve. Taylors Theorem is used to create a linear approximation to the non-linear statistic. The variance of a linear function is much easier to calculate and so the variance of the approximation is used to approximate the actual variance we are interested in. See Woodruff (1971) and Wolter (1985) for further details about how this method is derived.

For a stratified cluster design the method of approximation is applied to cluster totals within the stratum and the stratum totals are then summed. The formulas for the Taylor Series approximation, as implemented in SAS, STATA and SPSS, when the variable of interest is a mean, are of the following form:

Firstly, the weighted mean of the variable of interest is denoted as:

$$\overline{x} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} x_{hij} / W \tag{2}$$

Where:

$h=1,2,\dots,H$ is the stratum number, with a total of H strata

$i=1,2,\dots,n_h$ is the cluster number with stratum h, with a total of n_h clusters

$j=1,2,\dots,m_{hi}$ is the respondent number within cluster i of stratum h, with total of m_{hi} units

w_{hij} denotes the sampling weight for respondent j in cluster i of stratum h

x_{hij} denotes the observed values of the analysis variables for respondent j in cluster i of stratum h and

$$W = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \tag{3}$$

is the sum of all weights across all groups.

[1] See http://www.melbourneinstitute.com/hilda/doc/doc_hildamanual.htm

The standard error of the mean is calculated as the square root of the sum of all stratum variances:

$$stderr = \sqrt{V\left(\overline{x}\right)} \tag{4}$$

Where:

$$V\left(\overline{x}\right) = \sum_{h=1}^{H} V_h\left(\overline{x}\right) \tag{5}$$

is the sum of the variance from all stratum components.

Each stratum variance is denoted as:

$$V_h\left(\overline{x}\right) = \frac{n_h\left(1-f_h\right)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi.}-\overline{e}_{h..})^2 \tag{6}$$

Where:

f is the finite population correction – approximately zero in this situation.

$$e_{hi.} = \left(\sum_{j=1}^{n_h} w_{hij}\left(y_{hij}-\overline{x}\right)\right)/W \tag{7}$$

is the weighted deviation of each unit within cluster *i* of stratum *h* from the overall mean.

$$\overline{e}_{h..} = \left(\sum_{i=1}^{n_h} e_{hi.}\right)/n_h \tag{8}$$

is the average ‘deviation from the overall mean’ across all clusters within stratum *h*.

Delete-a-Group Jackknife

The estimator of the variance of $\hat{x}$ for the delete-a-group jackknife method is obtained by first constructing a set of R replicates weights $w_i^{(r)}$ for r = 1, …, R. For each set of replicate weights, an estimator $\hat{x}^{(r)}$ is then computed in the same way that $\hat{x}$ is computed using the main weights w_i. An estimator of the variance, and standard error, of $\hat{x}$ is then given by

$$V\left(\hat{x}\right) = \frac{R-1}{R}\sum_{r=1}^{R}(\hat{x}^{(r)}-\hat{x})^2 \tag{9}$$

$$stderr = \sqrt{V\left(\overline{x}\right)} \tag{10}$$

$\hat{x}$ can be a mean, population total or estimated model coefficient and the same formula for variance applies.

The construction of the replicate weights $w_i^{(r)}$ involves firstly creating R different replicate groups. The theory behind the jackknife method requires the replicate groups to consist of a subset of the data that mirrors the sample design properties of the full dataset. The HILDA sample was selected systematically from an ordered list of clusters so, to reproduce this for a subset, each cluster selected for the main sample is systematically allocated to a replicate group, with the process repeating once the first R clusters have been allocated to each replicate 1,…,R. All clusters allocated to a replicate are ‘dropped’ from that group and are excluded from receiving a weight. Each replicate group represents a smaller sample (1/R smaller) of units compared to the main sample but, due to the systematic dropping of clusters, retains a similar sample design.[2] A group of clusters were dropped from each replicate group, hence the delete-a-group jackknife method (Kott 1998).

Once replicate groups are formed the replicate weights are constructed by repeating the main weight creation process on each replicate group. The initial selection weights are allocated to each replicate group to form initial replicate weights, and then non-response adjustments and the benchmarking process are separately applied to each of these R sets of weights. The variability across the final replicate weights $w_i^{(r)}$, for a particular estimate, reflects the inherent variability of the originally selected sample.

The HILDA dataset provides 45 replicate weights (R=45) for each type of weight on the datasets.

[2] The ‘distance’ between clusters selected in a replicate is larger where a cluster has been dropped. This means that the sample design is not accurately reflected in each replicate group but it should not have a large impact on the results.

Data

For a complete summary of the HILDA data refer to the HILDA User Manual. For convenience the variables required in the construction of standard errors are briefly presented here. Table 1 outlines the variable names for the weights and replicates in the HILDA datasets. The "_" is used to denote the wave the variable is referring to. Replace with *a* for wave 1, *b* for wave 2, *c* for wave 3 etc.

Table 1: Weights

Population	*Weight*	*Replicate Weights*
Responding People (longitudinal)	_lnwtrp	_rwlnr1 - _rwlnr45
Enumerated People (longitudinal)	_lnwte	_rwlne1 - _rwlne45
Households (cross-sectional)	_hhwth	_rwh1 - _rwh45
Responding People (cross-sectional)	_hhwtrp	_rwrp1 - _rwrp45
Enumerated People (cross-sectional)	_hhwte	_rwlne1 - _rwlne45

Examples given later in this user guide will use "_wgt' and "_repwgt" to show where the code requires a variable from Table 1.

The longitudinal weights referred to in Table 1 are balanced panel[3] weights from wave 1 to the wave they refer to. For example, dlnwte is the balanced longitudinal enumerated person weight from wave 1 to wave 4. The Release 6 HILDA data has an additional data file (*longitudinal_weights_f60c*) containing all combinations of balanced panel and paired[4] longitudinal weights for both enumerated and responding persons (including longitudinal weight panels starting after wave 1). Replicate weights are not provided with the additional longitudinal weights but can be supplied on request (email hilda-inquiries@unimelb.edu.au).

In addition to weights, the sample design characteristics are required to describe the HILDA Survey design and need to be specified for any estimate via the Taylor Series Linearisation method.

Table 2: Sample Design Characteristics

Characteristic	*Variable*
Stratum	_hhstrat
Cluster	_hhraid

[3] A balanced panel of respondents includes only individuals that were responding or out of scope (dead or overseas) in all relevant waves. A balanced panel of enumerated persons requires the individuals to be enumerated or out of scope in all relevant waves.

[4] Paired responding person longitudinal weights include only individuals that were responding or out of scope (dead or overseas) in the first and last wave of the panel. Paired enumerated person longitudinal weights include only individuals enumerated in the first and last wave of the panel.

Software

This section outlines what the statistical packages STATA, SPSS and SAS provide in the area of variance calculation for complex surveys. A very useful summary of methods, of which some of the details in this section are based, is the DACSEIS evaluation project (http://www.dacseis.de/). An example of the code used in each program is provided and should be readily adaptable to other estimates of interest. For further information, consult the help file for the relevant program.

STATA

STATA has procedures available that produce estimates by the Taylor Series linearization method or incorporate jackknife replicate weights. The different characteristics of the sampling design, or the replicate weights, can be defined in STATA using a specific command. After the sample design characteristics are set, STATA automatically accounts for these sampling features in the survey commands (given the correct reference is used in the program). The use of replicate weights is a new feature of STATA version 9 and the commands specific to this will not work with prior versions of the program.

Examples

Specify the survey design characteristics of the HILDA dataset so that Taylor Series linearisation standard errors can be are produced. Once the survey characteristics have been specified any request prefixed but the 'svy:' term will produce standard errors via the correct method.

```
svyset [pweight=_wgt], strata(_hhmsr) psu(_hhraid)
```

Produce a mean (for income), with the correct standard errors:

```
svy: mean income
```

To break the mean down into subgroups (agegroup in this example) add the following to the line of the command above. (The subpopulation must be coded as a dummy variable (either a 0 or 1).)

```
svy: mean income, subpop(agegroup)
```

The next line of code runs a simple regression example, while producing the correct standard errors, using **income** as the response (dependent) variable and **agegroup** as the predictor (independent) variable.

```
svy: regress income agegroup
```

To automatically create standard errors via the Jackknife method the following code can be used to specify the characteristics required. The multiplier is simply 44/45 and represents the first term in equation (9). Similar to the Taylor Series method, once the survey characteristics have been specified any request prefixed but the ‘svy jackknife:’ term will produce standard errors via the Jackknife method.

```
Svyset [pweight =_wgt], jkrweight(_repwgt1 - _repwt45, multiplier(0.966667))
vce(jackknife) mse
```

The line below runs a simple regression example using **income** as the response (dependent) variable and **agegroup** as the predictor (independent) variable. The command will run (and run faster) without the jackknife option after the **svy** but you will get linearized standard errors instead of the jack-knife standard error.

```
svy jackknife: regress income agegroup
```

Mean income with jack-knife standard errors:

```
svy jackknife: mean income
```

SPSS

SPSS Release 12 has the add-on module ”SPSS Complex Samples”, which includes a set of procedures offering the possibility to analyse a complex sample:

- Complex Samples Plan (CSPLAN) for specifying a sampling scheme and defining the plan file used by the following procedures
- Complex Samples Descriptives (CSDESCRIPTIVES) for estimating means, sums and ratios of variables, computing standard errors, design effects, confidence intervals and hypothesis tests for the drawn sample
- Complex Samples Tabulate (CSTABULATE) for displaying one-way frequency tables or two-way cross-tabulations, related to the above-mentioned descriptive statistics which can be requested by subgroups.
- Complex Samples GLM (CSGLM) for running linear regression analysis, and analysis of variance and covariance.
- Complex Samples Logistic (CSLOGISTIC) for running logistic regression analysis on a binary or multinomial dependent variable using the generalized link function.
- Complex Samples Ordinal (CSORDINAL) for fitting a cumulative odds model to an ordinal dependent variable for data that have been collected according to a complex sampling design.

Once specified the module produces standard errors via the Taylor Series Linearisation method. Currently there is no module available in SPSS to automatically handle replicate weights, although the SPSS programming language allows for this to be manually written. The code examples provided below can all be replicated but the SPSS point and click interface.

Examples

Specify the sample design characteristics (this step is required for any of the later examples to work).

```
CSPLAN analysis
/plan file=’filename’
/planvars analysisweight=_wgt
/design strata=_hhmsr cluster=_hhraid
/estimator type=wr.
```

Produce an estimate of the total population mean, with Taylor Series standard errors:

```
CSDESCRIPTIVES
/plan file=’filename’
/summary variables=income
/mean.
```

To break the mean down into subgroups (agegroup in this example) add the following line to the procedure:

```
/subpop table=agegroup.
```

Produce a frequency table, with correct standard errors, of the number of people with a university degree (given a user derived ‘uni’ dummy variable):

```
CSTABULATE
/plan file=’filename’
/tables variables=uni.
```

The same subpopulation line, as given after the CSDESCRIPTIVES example, can be applied to the CSTABULATE command.

Run a regression, with correct standard errors, modeling income by age and sex.

```
CSGLM tifei by ahgsex with ahgage
/PLAN FILE=DataSet2
/MODEL ahgage ahgsex
/STATISTICS PARAMETER SE CINTERVAL.
```

SAS

There are four procedures for calculating results for complex surveys in SAS: SURVEYMEANS, SURVEYREG, SURVEYFREQ and SURVEYLOGISTIC (the last two only in version 9 onwards). The SURVEYMEANS procedure produces estimates of survey population totals and means with standard error and confidence interval calculation. SURVEYREG performs regression analysis for sample survey data, fitting linear models and computing regression coefficients and the covariance matrix. For one to n-way frequency and crosstabulation tables there is the SURVEYFREQ procedure, with corresponding variance estimation methods. It also allows design-based tests of association between variables, and for 2x2 tables the risk differences, odds ratios, relative risks and their confidence limits. SURVEYLOGISTIC performs logistic regression for categorical responses in sample survey data.

The built in procedures take both stratification and clusters into account and produce variance estimates via the Taylor Series linearization method. There is currently no built in procedures to automatically handle replicate weights in SAS. Outside of the available procedures the SAS programming language provides possibilities for all kinds of specific methods concerning variance estimation. The GREGWT weighting package, produced by the Australian Bureau of Statistics (ABS), contains a macro that handles replicate weights. An example invoking the TABLE macro, provided with GREGWT, and code to produce a similar result is given below.

Examples

Produce an estimate of population mean of income, with Taylor Series standard errors:

```
proc surveymeans data=w1r;
stratum _hhmsr;
cluster _hhraid;
weight _wgt;
var income;
run;
```

Various options are available for changing the measurement that is being estimated and the statistics output from the procedure. Consult the SAS help file for further options. To break the mean down into subgroups (agegroup in this example) change the procedure to the following:

```
proc surveymeans data=w1r;
stratum _hhmsr;
cluster _hhraid;
weight _wgt;
var income;
domain agegroup;
run;
```

Produce an estimate of the population total, with errors adjusted for the complex survey, of the number of people with and without a university degree (user derived ‘uni’ dummy variable).

```
proc surveyfreq data=w1r;
stratum _hhmsr;
cluster _hhraid;
weight _wgt;
table uni;
run;
```

Run a regression, with errors adjusted for the complex survey, modeling income by sex, age and university degree.

```
proc surveyreg data=w1r;
stratum _hhmsr;
cluster _hhraid;
weight _wgt;
model income = sex age uni;
run;
```

Run a logistic regression, with errors adjusted for the complex survey, modeling the probability of a university degree by sex, age and income.

```
proc surveylogistic data=w1r;
stratum _hhmsr;
cluster _hhraid;
weight _wgt;
model uni = sex age income;
run;
```

An alternative to the Taylor Series approach is to use the replicate weights. One way to implement them, to calculate mean income with correct standard errors is to run the SAS TABLE macro provided with the GREGWT weighting package:

```
%table(data=input, weight=_wgt, repwts=_repwgt1 - _repwt45, out=table1, var=income,
denom=_one_);
```

Code for the manual calculation of mean income, with standard errors produced via the replicate weights:

```
proc summary data=input;
  var income;
  weight _wgt;
  output out=mean0 mean=est0;
run;
%macro means;
%do i = 1 %to 45;
proc summary data=w1r;
  var income;
  weight _repwgt &i.;
  output out=mean&i. mean=est&i.;
run;
%end;
%mend means;
%means;
data samerr (keep = est0 stddev samerr VarEst);
      merge mean0 mean1 mean2 mean3 mean4 mean5 mean6 mean7 mean8 mean9
mean10 mean11 mean12 mean13 mean14 mean15 mean16 mean17 mean18 mean19
mean20 mean21 mean22 mean23 mean24 mean25 mean26 mean27 mean28 mean29
mean30 mean31 mean32 mean33 mean34 mean35 mean36 mean37 mean38 mean39
mean40 mean41 mean42 mean43 mean44 mean45;
      by _type_ _freq_;
```

```
     array est [46] est0-est45;
     array sqdiff [45] sqdiff1-sqdiff45;
     do i=1 to 45;
          sqdiff[i]=(est[i] - est0)**2;
     end;
     VarEst=44/45* sum(of sqdiff1-sqdiff45);
     stddev=sqrt(VarEst);
     samerr=1.96*stddev;
run;
```

More complex SAS programs can be written to create the correct standard errors, using replicate weights, for regression procedures.

References

DACSEIS Project http://www.dacseis.de/

Kott, P.S. (1998), 'Using the Delete-a-Group Jackknife Variance estimator in Practice', *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 763-768.

Machlin, S., Yu, W., and Zodet, M. (2005), ‘Computing Standard Errors for MEPS Estimates’, Agency for Healthcare Research and Quality, Rockville, Md. http://www.meps.ahrq.gov/data_stats/standard_errors.jsp

STATA library http://www.ats.ucla.edu/stat/stata/Library/replicate_weights.htm

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

Woodruff, R.S. (1971), ‘A Simple Method for Approximating the Variance of a Complicated Estimate’, *Journal of the American Statistical Association*, 66, 411 - 414.