Calculating Design-Corrected Standard Errors for the General Social Survey, 1975-2012

Steven Pedlow, NORC at the University of Chicago (pedlow-steven@norc.uchicago.edu)

Revised – June 7, 2013

The General Social Survey (GSS) is an area-probability sample that uses the NORC National Sampling Frame for an equal-probability multi-stage cluster sample of housing units for the entire United States. Since the sample for the GSS is a cluster sample, standard errors are larger for the GSS than simple random sample calculations (calculated without correction for the design). To correctly calculate standard errors, design variables must be used in statistical software (such as PROC SURVEYFREQ in SAS). Without these design variables, statistical software will assume a simple random sample and underestimate standard errors.

We provide two design variables for every GSS interview from 1975-2012, VSTRAT and VPSU. VSTRAT is the Variance STRATum while VPSU is the Variance Primary Sampling Unit. The stratum and PSU reflect the first-stage and second-stage units selected as part of the NORC National Sampling Frame, and are unique to a particular round. There are two second-stage units (VPSU) for each first-stage unit (VSTRAT).

First stage units in the NORC National Sampling Frame are called National Frame Areas, (NFAs), each of which is composed of one or more counties (previous to the 2010 National Frame, NFAs were called PSUs). The largest urban areas are selected with certainty to guarantee their representation in NORC's National Sampling Frame. Second-stage stage units in the NORC National Sampling Frame are called segments, each of which is either a block, a group of blocks, or an entire census tract. The first-stage and second-stage units are selected with probabilities proportional to size (in housing units), and the sample housing units (third-stage units) are then selected to be an equal-probability sample, which results in roughly the same number of housing units selected per second-stage sampling unit.

To create the variables VSTRAT and VPSU, we recode the NFAs and segments, depending on whether the NFA was selected with certainty. In certainty NFAs, segments are paired into strata with one segment assigned to VPSU = 1 while the other segment is assigned to VPSU = 2. Often, small segments are combined into one VPSU. Non-certainty NFAs are paired into strata with one NFA assigned to VPSU = 1 while the other NFA is assigned to VPSU = 2. It is rare, but possible, for NFAs to be combined in one VPSU. This strategy has been adapted from the National Longitudinal Survey of Youth, 1997 cohort strategy designed by Kirk Wolter.

Here is sample Stata code to analyze the variable ANALYSISVAR within a GSSDATAFILE with the weight variable WTVAR (either WTSSALL or WTSSNR):

```
use GSSDATAFILE.dta, clear
svyset vpsu [weight=WTVAR], strata (vstrat)
svy: proportion ANALYSISVAR // point estimates and design adjusted s.e.'s
svy: tabulate ANALYSISVAR, deff //deff
tab ANALYSISVAR [aweight= WTVAR],missing // Weighted frequency
```

Note that it is possible to combine multiple years of GSS data into one GSSDATAFILE. SPSS is menu-driven, so no code is given here, but you can create design-corrected standard errors within SPSS using the Complex Samples add-on.

STATA error handling: "missing standard error because of stratum with single sampling unit"

VSTRAT and VPSU were created so that there was a minimum of 3 GSS respondents within a VSTRAT/VPSU cell. If all three are missing on a variable, this error can occur in Stata. If a GSS round is subset (to males or females, for example), this error becomes more likely to happen.

The best workaround is to merge two VSTRATA together to eliminate this problem (the VSTRATA are ordered so that similar VSTRATA are numerically consecutive). To diagnose the problem, run a frequency of the data by VSTRATA/VPSU and look for VSTRAT values with only one VPSU with respondents. Here is an example:

VSTRATA	**VPSU**	**# of cases**
…	…	…
x-1	1	3
x-1	2	5
x	1	4
x+1	1	6
x+1	2	4
…	…	…

The error occurs because VSTRATA = x has four cases with VPSU=1 but none with VPSU=2. This prevents Stata from calculating the variance for this strata (it has nothing to compare with VPSU=1). One solution is to combine VSTRATA x-1 and x (if VPSU =2 for VSTRATA=x, combining VSTRATA x and x+1 could be done) in two steps:

1. If VSTRATA = x-1 and VPSU = 2 then VPSU=1
2. If VSTRATA = x then VSTRATA=x-1 and VPSU=2.

Here is the revised frequency:

VSTRATA	**VPSU**	**# of cases**
…	…	…
x-1	1	3
x-1	~~2~~ 1	5
~~x~~ x-1	~~1~~ 2	4
x+1	1	6
x+1	2	4
…	…	…

This eliminates the "stratum with a single sampling unit". In severe cases of data subsets, this step may be required more than once.