Stratum and Cluster Variables for SDA Version of GSS

The GSS file distributed by NORC now has variables named VSTRAT and VPSU which are intended to be used as stratum and cluster variables for the purpose of calculating standard errors. The VSTRAT variable, however, is unusable by SDA as a stratum variable for calculating standard errors because it has thousands of categories. The VPSU variable, which was designed to be used together with VSTRAT to identify the clusters within each stratum, is therefore also unusable.

Note, however, that since the individual strata identified by VSTRAT are all substrata of the various strata identified by SDASTRATA (used by SDA, as described next), the calculated standard errors should be very similar.

Stratum Variable (SDASTRATA)

The GSS has a variable named SAMPLE, which identifies the sampling frame used by NORC and the GSS over the decades. That sampling frame was updated every ten years, in line with the decennial census. The first category of SAMPLE was for the 1960 sampling frame, which was still used in 1972, for the first year of the GSS. Additional categories of that variable correspond to subsequent updates of the NORC sampling frame.

The stratum variable used by SDA is a recoded version of SAMPLE named SDASTRATA. The original GSS SAMPLE variable is recoded to combine category 3 with 4 and category 6 with 7, so that oversampled black cases (in 1982 and 1987) in a specific geographic area are in the same SDASTRATA category as the cases in the regular sample in the same geographic area. Since the number of black oversampled cases in most of the geographic areas (identified by SAMPCODE) was very small, those cases would otherwise be excluded from many calculations that required complex standard errors.

Finally, a new category was added to SDASTRATA for the 2021 GSS. Whereas the GSS sample in the previous years used the NORC sampling frame for in-person interviews based on cluster sampling, the 2021 GSS was primarily a Web-based study because of the Covid-19 pandemic. The 2021 sample was mostly an un-clustered Address Based Sample (ABS) of U.S. households. Since it was based on a new sampling frame, it was identified by a new category in SDASTRATA – the variable used by SDA as the stratum variable.

Cluster Variable (SDACLUSTERS)

The GSS has a variable named SAMPCODE which identifies the areas selected into the sample. For most years, this variable serves as the cluster variable, but there are some years for which this does not apply. Depending on the year, the cluster variable for SDA (SDACLUSTERS) was created as follows:

- The 1972 study (the first year of the GSS) was a block-quota sample based on the 1960 NORC sampling frame. The clusters were not identified by a code in the SAMPCODE variable. Therefore, the 1972 cases would not be included in the calculation of complex standard errors unless we provided pseudo-clusters for that year. Without this procedure, the 1972 cases were considered as all coming from the same PSU in a single stratum, and they were therefore dropped from the calculations whenever the design variables (the stratum and cluster variables) were used to calculate standard errors or summary statistics. Since the average size of the clusters identified by SAMPCODE was about 100 cases, we divided the 1,603 cases from 1972 into 16 random groups to create the pseudo-clusters for the SDACLUSTERS variable.

- The studies in 1973-2018 were typical cluster samples, and the SAMPCODE variable was simply copied over to the SDACLUSTERS variable for those years.

- The 2021 GSS was a completely new and different sample. About a fifth of the sample (6000 addresses) was a multi-stage sample selected from the 2010 NORC frame, intended originally for in-person interviews. Because of the pandemic, however, the fieldwork plans were changed to primarily a Web-based study. That original sample was supplemented by the addition of about 20,000 addresses selected as an un-clustered address-based sample. Most of the 2021 cases, therefore, are un-clustered, and the calculation of standard errors for 2021 alone could be carried out in a different way. However, since SDA usually calculates complex standard errors based on the cases from all the years combined as a single stratified cluster sample, we needed to provide the 2021 sample with pseudo-clusters (as for 1972). Since there were 4,032 cases from 2021, and since we wanted clusters of about 100 cases, we divided the 4,032 cases from 2021 into 40 random groups to create the pseudo-clusters for the SDACLUSTERS variable.

See the section on “Sampling Design and Weights” in the full NORC codebook for the 2021 GSS for more information on sampling and weighting.

For more information on how SDA uses the stratum and cluster variables to calculate complex standard errors, see the online help file at: https://sda.berkeley.edu/sdaweb/helpfiles/semethod.htm