SUBSAMPLE REPLICATION APPLIED TO DATA FROM THE GENERAL SOCIAL SURVEYS

Neil W. Henry and Ping Yin Kuan, Virginia Commonwealth University

Despite increased emphasis on quantitative analysis in sociology over the past two decades the average level of statistical competence on the part of social science researchers remains remarkably low. These years have seen the successive adoption of interesting statistical methods such as path analysis, regression analysis, confirmatory factor analysis, and loglinear modelling by researchers who overestimate the power of these methods to overcome serious methodological problems in their data. This results in misuse and misinterpretations in articles even in the most prestigious journals in the discipline.

In this paper we consider the problem of the analysis of data from a complex sample survey. Although simple random sampling of human populations is relatively rare in social research, students are not usually exposed in their statistics courses to discussion of the problems of making inferences from non-SRS samples. Textbooks, after a brief description of the mechanical aspects of cluster sampling, typically include a nominal reference to Kish (1965), but there is little evidence of in-depth study. Blalock (1979:569) is exceptional in including a formula for the relative efficiency of SRS and a cluster design, but only for estimation of a mean. The relevance of the discussion to the multivariate analyses that are so commonplace nowadays is not apparent.

The major research centers that carry out sample surveys have,of course, been well served by survey statisticians. Primary analysis of large scale surveys carried out under their auspices reflect the most sophisticated and up to date thinking in both the design and analysis of survey data. The problem of inferior analysis is most often seen in secondary analyses of complex data sets, carried out in venues far removed from their places of origin.

A prime recipient of this sort of misuse is the General Social Survey carried out by the National Opinion Research Center since 1972. This survey of some 1500 Americans explores opinions and attitudes on a wide range of current issues, while also collecting demographic information. It thus is an attractive data source for social scientists of a variety of theoretical persuasions.Copies of the dataset generated by the surveys have been widely distributed throughout the country in machine readable form, and are available to anyone who has access to a good-sized university computer. Its availability has resulted in its wide use by sociologists as a teaching tool and a research resource.

The GSS uses a multi-stage areal cluster design that is self-weighting, with some stratification based on census data . The first stage consisted of selection of 101 counties or SMSAs from the list of all such units in the country.These units were ordered by region of the country and by racial composition, so that some stratification on these variables occurred. The primary sampling units were selected with probabilities based on their 1970 Census populations, so that "self-weighting" ignores population shifts since 1970. It is important to note that the same 101 PSUs (which include 89 distinct SMSAs and counties) are used for each of the yearly surveys.

At the second stage Census-defined block groups (or enumeration districts where appropriate) were selected, again with probabilities proportional to their size after an ordering that incorporated geographical location, race, and income information. Next, blocks and finally households within blocks were selected, with an average of five persons per block interviewed in each survey.

The information about GSS design given above was drawn from the public documentation (Davis,1980:Appendix A) that is available to all potential users of the dataset. There is no reason to doubt the claim, found therein, that:

"The NORC national probability frame, with its broad geographic dispersion...provides sufficient flexibility for application to a wide range of survey tasks."

The question that needs to be addressed, however, is whether the average user of the dataset has the knowledge and the resources to carry out appropriate statistical analyses of the data.

What kinds of analysis are easy to carry out using this particular data set? Simple percentages and means will

provide unbiased estimates of their population equivalents, due to the self-weighting nature of the design. Ordinary regression analysis will also give estimates that consistently estimate population parameters. Simple unweighted variances and standard deviations calculated from the sample according to the usual formulas will be biased because of the clustering process, however. As a result, correlation coefficients and standardized regression coefficients ("beta weights" in the usual terminology) calculated from the sample also have an ambiguous relationship to their population counterparts. Of course population variances, correlations and the like can be estimated consistently by using information about the degree of intrablock in contrast to interblock variability. Appropriate algorithms are not commonly used by the typical user of the GSS, however.One reason for this is that there is no discussion of the problem in the documentation that accompanies the dataset and no methods specified in the computer packages he or she is likely to call upon in the course of a data analysis. (Another reason is that editors and reviewers seem uncritical of analyses totally based on SRS assumptions.)

Furthermore, tests of significance and confidence coefficients computed on the basis of simple random sampling assumptions will be in error, since they rely on the availability of unbiased estimates of variance. While social scientists are notoriously cavalier in their attitude toward inferential statistics (perhaps because they never have to pay for their errors, of whatever Type), they continue to publish p-values and t-values in their research, and use them in their arguments.

One approach that is commonly used in analyzing data from complex samples is to make a blanket adjustment to the sample size in order to take into account the fact that the attained sample is less efficient than a simple random sample of the same size. Usually the primary researcher, knowledgeable about the design, will suggest such an adjustment factor after some preliminary analyses of the data. The GSS documentation suggests that the yearly samples of approximately 1500 persons correspond to "effective" simple random samples of size 1000 (Davis, 1980: 187). Even this crude adjustment factor is ignored in the vast majority of articles published in sociological journals that use these surveys.

The work of Frankel (1971) and Kish and Frankel (1970), however, implies that a single efficiency factor is not sufficient when regression and correlation studies are being done. Frankel's study found design effects of 22 percent for simple correlations and 50 percent for multiple correlations in a situation where the design effect in estimation of means was 30 percent. The effect of the design on regression coefficients was less, only about 10% (Frankel, 1971:116).

We have chosen to use simple subsample replication to estimate parameters and their standard errors in a typical use of the GSS data: regression of respondent's income on his education,on aspects of his occupation, and on his father's education and occupation. The SAS package is easily utilized for the task. Sudman (1976:174) provides an excellent presentation of the subsample replication method, in a context that is practically identical to the design of the GSS sample. When a sample can be broken up into k statistically identical subsamples, the variation in the k subsample estimates of a parameter can be used to estimate directly the standard error of the estimate calculated from the complete sample. Let θ be the parameter in question, $\hat{\theta}$ its full-sample unbiased estimator, and $\hat{\theta}_i$, i = 1 to k, the k subsample estimates. Then the standard error of $\hat{\theta}$ is estimated as follows:

$$\widehat{SE(\hat{\theta})} = SQRT[\Sigma (\hat{\theta}_i - \bar{\theta})^2/k(k-1)]$$

where

$$\bar{\theta} = \Sigma \hat{\theta}_i /k$$

In the GSS we can form statistically equivalent subsamples by sorting on the 101 PSUs. This information is available in the data file, and the PSUs are ordered in the same way that they were originally selected for the sample. Thus, by taking the odd- numbered PSUs, for example, we have the sample that would have been gathered if the decision had been made to collect only about 750 interviews instead of 1500. In our analysis 10 subsamples were formed by combining every 10th PSU. Sudman notes that the choice of k is problematic; more than 10 would leave fairly small subsamples, while the minimum of 2 provides less information about sample to sample variability.

The ambiguous phrase "statistically equivalent subsamples" was used above, and needs some clarification. If we form 10 subsamples from this particular

data set by combining data from every tenth PSU into a subsample, the subsamples will not contain the same number of cases, nor even the same number of PSUs, since 101 is not evenly divisible by 10. In applying the formula above ,however, we are assuming that each subsample estimate, $\hat{\theta}_i$, conveys the same amount of information about the parameter θ. While this is not the case, we will argue that the subsample to subsample variation is more crucial to the estimation of standard errors than the exact number of observations within each subsample. This latter number should not vary a great deal, however, if the method is to be reasonably accurate.

The calculations needed to estimate the standard errors of estimated regression coefficients are easily ordered using SAS; so easily, in fact, that one might call this a trivial amendment of the usual program. Suppose that the variables Y, X1, X2, X3, X4, and INDEX are contained in a SAS dataset called DATA1. Y will be the dependent variable in the regression, the Xs the independent variables, and INDEX indexes the subsamples, taking on integer values from 1 to k. In the case of the GSS, SAMPCODE is the name of the variable that identifies the 101 PSUs, and INDEX = MOD(SAMPCODE,10) defines the subsample index.

The SAS statements for an ordinary regression are :

```
PROC REG DATA = DATA1;
     MODEL Y = X1 X2 X3 X4;
```

For subsample estimation, they are:

```
PROC SORT; BY INDEX;
PROC REG OUTEST = COEFFS;
     MODEL Y = X1 X2 X3 X4/NOPRINT;
     BY INDEX;
PROC MEANS DATA = COEFFS MEAN
                  STDERR;
```

To simplify the printed output, only the information that is needed for this discussion has been requested. The ordinary output of PROC REG has been suppressed by NOPRINT. The various regression coefficients from the 10 subsample regressions are stored in the dataset COEFFS, and are averaged by PROC MEANS. Printed out by PROC MEANS are the averaged regression coefficients for each independent variable and the constant term under the heading 'MEAN' , and the estimated standard errors under the heading 'STDERR'.

Shown in Table 1 are the results of using this method to estimate an equation in which personal income is regressed on seven other variables related to education and occupation. The subsample analyzed consisted of adult males employed full-time outside the agricultural sector at the time of the survey. Data from the 1976 and 1977 surveys were combined for this analysis. Requiring that complete data be available on all variables in the equation resulted in a total sample size of 506. These results are drawn from Kuan (1982), where a multi-equation model was considered, and where a discussion of the theoretical basis for the analysis may be found.

Income is measured in thousands of (1976) dollars, education in years of schooling, and the prestige of an occupation using the 100 point Duncan scale. A dummy variable distinguishes white-collar from blue-collar jobs, while "authority level" is a six-point index constructed by Kuan (1982) which indiicates the extent to which the individual supervises, or is supervised by, others on the job.

It was expected that the standard errors estimated using subsample replication would be larger than those produced by the usual SRS calculations on the full sample. Of the eight standard errors reported in Table 1, five showed the opposite relationship, the SSR estimate being smaller than its SRS counterpart. This is not an isolated instance: in the larger study about one-third of the standard errors of regression coefficients showed this pattern. The findings are consistent with Frankel's (1971) conclusion that SRS estimates of standard errors of regression coefficients appeared to be less biased than those of means, and suggests that the efficiency adjustment proposed in the GSS documentation may be overly conservative as far as unstandardized coefficients are concerned.

The simplicity of the subsample replication method, and the fact that it can be implemented using a widely available package such as SAS, should be brought to the attention of users of secondary data such as the GSS, and should also be included in middle level statistics courses for social scientists.

Table 1

Regression of Income on Other Attributes

Employed White Males, N = 506

Independent Variable	Regression Coefficient		Estimated Standard Error	
	$\hat{\beta}$	$\bar{\beta}$	SSR	SRS
Constant	6.111	5.887	2.559	2.002
Education(F)	-.051	-.003	.078	.104
White collar(F)	.369	.181	.678	.956
Prestige(F)	.015	.012	.044	.035
Education	.480	.494	.147	.168
White Collar	1.487	1.527	.899	.983
Prestige	.045	.035	.030	.034
Authority	.952	.978	.291	.236

Notes:

(F)	Reported attribute of father
SSR	Subsample replication estimate
SRS	Full sample OLS estimate
$\bar{\beta}$	Average over 10 subsamples
$\hat{\beta}$	Full sample OLS estimate

REFERENCES

Blalock,Hubert M. 1979. SOCIAL STATISTICS. N.Y.:McGraw-Hill

Davis,James A. 1980. GENERAL SOCIAL SURVEYS, 1972-1980;CUMULATIVE CODEBOOK. Chicago: NORC.

Frankel,Martin R. 1971. INFERENCE FROM SURVEY SAMPLES. Ann Arbor: University of Michigan.

Kish,Lesliie 1965. SURVEY SAMPLING. N.Y.: John Wiley.

Kish,Leslie and M.R.Frankel 1970. Balanced repeated replications for standard errors. JASA, 65, 1071-1094.

Kuan,Ping Yin 1982.Objective Statuses and Class Identification: A Causal Analysis. Unpublished Master's Thesis, Virginia Commonwealth University.

Sudman,Seymour 1976. APPLIED SAMPLING. N.Y.: Academic Press.