SURVEY CALIBRATION TO CPS HOUSEHOLD STATISTICS

Varma Nadimpalli, David Judkins, and Adam Chu; Westat
Varma Nadimpalli, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Raking, Poststratification, Control Totals

1. Introduction

It is common practice in many household surveys to poststratify the sampling weights to relevant population totals (i.e., control totals) derived from independent sources. The National Survey of Parents and Youth (NSPY), for example, uses a ratio-raking algorithm to calibrate the sampling weights to estimates of households with youth 9- to 18- years of age. The estimates used for this purpose are derived from the monthly Current Population Survey (CPS). The CPS control totals show quite a bit of month-to-month variation which can affect the levels of estimated totals derived from the survey. In addition, if the required CPS statistics are not yet available for a particular time period, outdated control totals may have to be used for weighting purposes.

The main focus of this paper is on the use of regression models to derive alternative estimates of household totals from available CPS household statistics. Such models can be used to "smooth" available CPS estimates as well as to extrapolate to time periods not covered by the existing data. To evaluate the method, we compare the original and smoothed control totals using the last 11 years of published CPS data.

The other focus of the paper is on ways of reflecting the variability of the CPS control totals on the corresponding estimates of sampling errors derived from the survey. For the NSPY, the replication method described in Rizzo and Judkins (2004) is used. To approximate the variance of the CPS totals in the estimates of variance derived from the NSPY, we used the standard errors obtained from the regression model to generate the replicate control totals needed to develop the replicate weights. Some limited evaluation of the use of the model-derived standard errors for this purpose, rather than those obtained from the generalized variance functions (GVFs) available in CPS reports, is provided.

2. Application

The National Youth Anti-Drug Media Campaign was funded by Congress to reduce and prevent drug use among young people 9- to 18- years of age, by addressing youth directly as well as indirectly, and by encouraging their parents and other adults to take actions known to affect youth drug use (Hornik, et.al., 2003). The primary tool for the evaluation is the National Survey of Parents and Youth (NSPY). The NSPY is a household-based survey with a sample of over 25,000 youth and 18,000 parents. Households were selected in stages using a stratified multistage probability sampling design. At the first stage of selection, 90 primary sampling units (PSUs) were selected from 50 strata with probabilities proportionate to size (PPS). The PSUs were generally metropolitan statistical areas (MSAs) or groups of non-metropolitan counties. Within the selected PSUs, a sample of 2,800 second-stage units referred to as segments was selected. The segments were of two types: area segments consisting of Census-defined blocks or block groups, and new construction segments consisting of groups of building permits issued by building permit offices over a specified interval of time. Within the sampled segments, 81,000 dwelling units were selected and screened to identify eligible households (i.e., households with youth 9- to 18- years of age). Over the course of the study, completed interviews were obtained for 25,000 youth and 18,000 parents associated with the sampled youth. For the NSPY, parents were defined to include natural parents, adoptive parents, foster parents who lived in the same household as the sampled youth, as well as stepparents and other relatives serving as parents provided they lived with the child for at least six months.

For analysis purposes, separate sets of sampling weights were developed for youth, parents, and youth-parent dyads (e.g. see Hornik, et.al., 2003, Appendix A), where a dyad was defined to be a unique youth-parent combination. All of the weights were designed to reflect overall selection probabilities and to compensate for nonresponse and undercoverage. Since only one parent was usually sampled per household while up to two youth could be sampled in the same household, a responding parent could be included in up to two distinct dyads. The weights for youth and dyads were developed using analogous procedures and involved a final poststratification (raking) step to CPS-based estimates of person-level population counts. The derivation of the parent weights, however, required an intermediate step involving the poststratification of the household weights to the corresponding CPS-based estimates of *household* counts. The goal of the raking was to reduce biases due to under-coverage and nonresponse, and to reduce the sampling error of the estimates. In the raking process, the weights were iteratively adjusted until the sum of the weights agreed with the corresponding population totals derived from the CPS.

For the first three waves of the NSPY (referred to as the 'recruitment" waves), the youth and dyad weights were raked to population counts (i.e., control totals) of youth 9- to 18- years of age. (For the subsequent follow-up waves, the weights were raked to counts of youth 12- to 18- years of age due to the aging of the sample.) The parent weights were not raked in the same fashion because no control totals exist for parents as defined for the NSPY. However,

estimates of total households with youth in the relevant age ranges are available from the CPS and were used to rake the household weights from which the parent weights were derived. In the remainder of this paper, we focus on the use of estimates of CPS household counts for weighting purposes.

3. Control Totals

The Current Population Survey (CPS) is often used to derive control totals for weighting purposes. Results from the Current Population Survey are published often and quickly. Moreover, it is relatively easy to tabulate CPS data from public use files, and the sample sizes are larger than those in the NSPY.

On the negative side, it is known that there are some problems in estimating the total number of households in the United States from the CPS. These problems occur primarily because of unknown apportionment of under-coverage between and within households. For more information about this problem, see Alexander (1987), Alexander (1990), and Alexander and Roebuck (1986). However, we felt that the advantages of using the CPS data far outweighed these problems.

The monthly estimates of household totals from the CPS normally show quite a bit of month-to-month variation due to sampling error. In addition we did find anomalous results for some months, apparently resulting from a technical problem that occurred in producing the public version of the CPS data files. Figure 1 shows the total number of households in the U.S. as derived from published CPS data files. We see a sharp decline in the household counts from May 1994 to May 1995. We determined that this was caused by corruption in the household ID. Because of this problem, it is not appropriate to use the household ID as an identifier to tabulate households for these months. We have notified the data publisher (National Bureau of Economic Research) about the problem. Users must exercise caution when using the public use data base to derive household counts. This problem does not exist for person-level estimates.

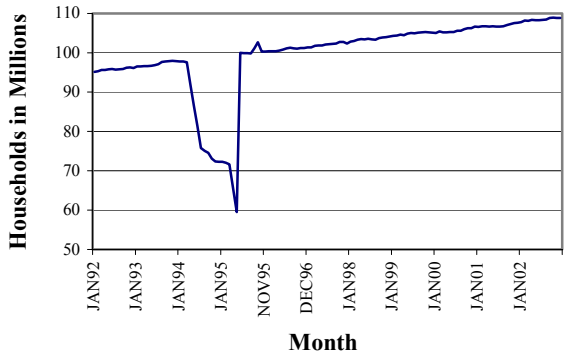

Figure 1. CPS estimates of total number of households in the United States

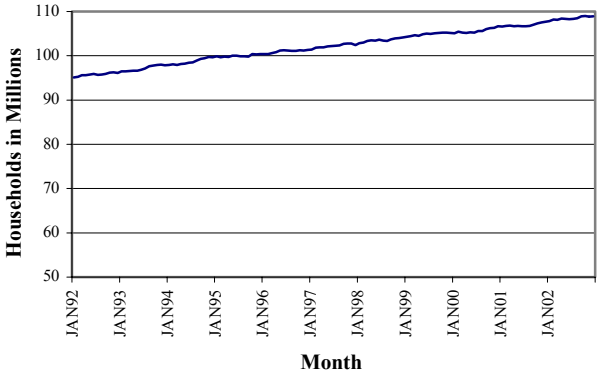

Figure 2. CPS household estimates using new household identifier

We were able to overcome the problem by treating the entire vector of household information in the CPS data files as a unique household identifier. Figure 2 gives the resulting household estimates using the new unique household "identifier." Figure 3 gives the corresponding number of households with youth 9- to 18- years of age using the new household identifier. The figure shows quite a bit of month-to-month variation due to sampling error. The degree of variation is not unexpected because the CPS household weights are not poststratified to independent control totals. We used the regular CPS weights and not the CPS March Supplement weights in these tabulations.

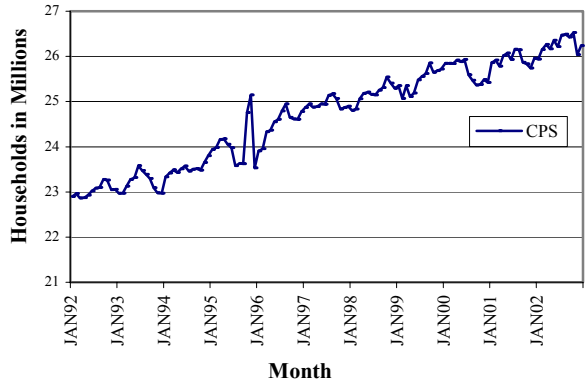

Figure 3. Households with youth 9 to 18 years of age

4. Smoothed Control Totals

It is desirable to smooth the CPS estimates to reduce the month-to-month variation in the control totals. Also, since the CPS estimates are themselves subject to sampling variability, we want to reflect that variability in the NSPY variance estimates (as discussed later in Section 6). The various options we considered for smoothing the control totals are:

1. Regression: To develop control totals for NSPY weighting, we used 21 months of CPS data. As

indicated below in Figure 4, the regression model produced estimates that are close to the CPS estimates.

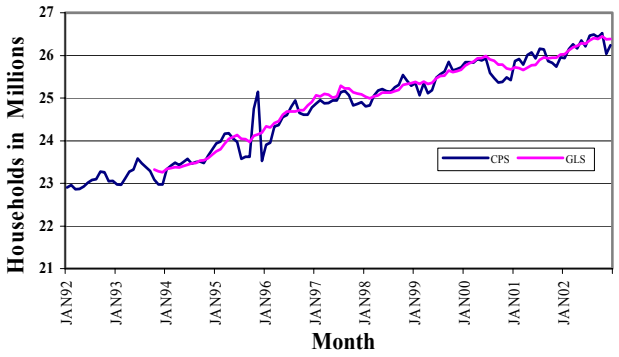

Figure 4. Comparison of CPS household estimates with estimates derived from regression model

2. Time Series: We worked with a method of moving averages, but the regression model produced better estimates with smaller standard errors. Weidman and Bobbitt (1991) have suggested smoothing household control totals using an ARIMA model with 72 months of data; however, we have not applied their method to these data.

The regression model we used to smooth the CPS-based control totals was the simple linear model given by:

$$y_i = \alpha + \beta i + \varepsilon_i, \quad i = 1,....,n \tag{1}$$

$$Cov\begin{pmatrix} \varepsilon_i \\ \vdots \\ \varepsilon_i + 20 \end{pmatrix} = V\sigma^2, Cov(\varepsilon_i, \varepsilon_i) = 0 \text{ for } j - i > 20$$

where y_i is the estimated number of households for month i based on the CPS, and V is a known and positive definite matrix. V incorporates the correlation structure of the monthly CPS estimates and was constructed using estimates of correlations presented in Gunlicks, Corteville, and Mansur (1987). Estimates of the parameters α and β in the model (1) were obtained using weighted least squares (WLS).

To evaluate the regression method, we secured 111 additional months of CPS data (for a total of 132 consecutive monthly estimates) so that we could repeat the smoothing process a total of 111 times. This was done as follows. Starting with i = 1, WLS estimates of the regression parameters were obtained using the first 21 data points (i.e., i = 1, 2, ..., 21). The fitted model was then used to estimate (predict) y_{22} (the next point in the data series). Using the model to predict a future value is consistent with the procedures used in the NSPY, where CPS control totals for the most recent time periods were not always available in time for weighting purposes. The same process was then applied to the 21 data points starting with i = 2 (i.e., i = 2, 3, ..., 22), and the resulting model was used to estimate the future value, y_{23}. The entire process was repeated for each successive group of 21 data points, resulting in 111 regression models and "future" predicted values.

Except for the months in which the decennial census is conducted, there are no "true" values with which to compare the CPS-based household counts. To obtain rough approximations of the true values, we used additional data points to predict the same values obtained from the 21-point regression models described above. That is, regression models of the form (1) were estimated for each successive group of 34 data points (again, for a total of 111 models) to predict the 22nd value in the respective series. The use of the additional data points was expected to provide more reliable estimates than the corresponding 21-point models because of the greater number of data points and the fact that the data points now span the time period of interest.

The predicted values from the 34-point model were used to compute the mean square error across the replications as follows. For month t, let

y_t = Monthly CPS estimate
$\tilde{y}_t$ = Smoothed estimate using 21 months
Y_t = Smoothed estimate with 21 prior + 12 subsequent months

Then the mean square errors of the smoothed and original CPS estimates respectively were computed as

$$MSE_{\text{smooth}} = \sum(\tilde{y}_t - Y_t)^2$$

and

$$MSE_{\text{cps}} = \sum(y_t - Y_t)^2$$

The reduction in the mean square error produced by smoothing was computed as:

$$1 - \frac{MSE_{\text{smooth}}}{MSE_{\text{cps}}} \tag{2}$$

5. Results of Smoothing

The overall reduction in mean square error due to using the regression-based estimates instead of the actual CPS estimates of the total number of households with youth 9- to 18- years of age was 65 percent. Reductions varied by domain. For example, the reduction in MSE for the Northeast region was 62 percent, compared with 98 percent for the Midwest, 54 percent for the South and 61 percent for the West.

We also used the regression model to estimate the total number of households for April 2000 and compared it

with the total number of households from the 2000 Census. According to the 2000 Census, the total number of households in the United States was 105.5 million. The corresponding estimate from the April 2000 CPS was 105.2 million, while the smoothed estimate derived from the regression function was 105.5 million.

6. Replicate Control Totals

The sampling variance of estimates derived from the NSPY can be obtained using a replication approach. The replication method developed for the NSPY is described in Rizzo and Judkins (2004). The method uses 100 replicates. This method reflects the between-stratum variance due to subsampling 90 PSUs from a larger sample consisting of 100 PSUs, and also reflects the finite population correction factors at both the PSU and segment levels.

It is important to reflect the variability of the CPS estimates in the construction of the replicate weights required for variance estimation. To do this, we poststratified (raked) the replicate weights to varying control totals that approximately reflected the CPS variance. The required replicated control totals were obtained using random numbers generated from a normal distribution with variance equal to the CPS variance. Judkins (1991) suggested using 'a' and 'b' parameters associated with published CPS generalized variance functions (GVFs) to obtain the variance of monthly CPS estimates. However, the GVFs give the approximate sampling variance of the CPS estimates and not the variance of the smoothed estimates. Since we used the predicted values from the regression model rather than the original CPS estimates, we decided to use the estimated variance of the predicted values based on the model to generate the replicate control totals. Thus, using the appropriate regression model, smoothed estimates of household counts and associated variances were obtained for every interior cell defined by the raking dimensions. We then summed the smoothed estimates corresponding to the interior cells to obtain the required marginal totals. Assuming that the CPS estimate at the level of the margins is approximately normally distributed, the replicate control total for a cell j and replicate k was derived as:

$$REP_{j,k} = Total_j + \left(\frac{SE_j * Z_{j,k}}{\sqrt{h_k} * \sqrt{100}} \right) \qquad (3)$$

where $Total_j$ is the smoothed control total for cell j, SE_j is the corresponding standard error obtained from the regression model, $Z_{j,k}$ is a standard normal random variable, and h_k is a scaling factor for replicate k (referred to as a JKN factor). The constant of 100 in formula (3) refers to the fact that 100 replicates are used for the NSPY.

The generalized variance functions (GVFs) that are provided with the CPS public use files are not always applicable to the control totals used for poststratification. For example, GVFs that apply to particular subsets of households with youth 9- to 18- years of age are generally not available. In such cases users may be forced to use the readily available (but possibly inappropriate) GVFs to obtain rough approximations of standard errors. Another application of the regression approach would therefore be to use the residual variance obtained from models to estimate the sampling variance of the original CPS estimates. Examples of such estimates (expressed as standard errors) are given in the first column of standard errors in Table 1. For comparison, we also computed the standard errors using the CPS GVF's available for March 2001. These are shown in the last column of Table 1. Note that while the GVF-based standard error corresponding to "total households" is close to the residual-based standard error, the standard errors corresponding to the 9- to 18-year old subgroups are not. This may be due to the fact that the GVFs used to obtain the subgroup standard errors are not appropriate. Finally, standard errors of the smoothed estimates obtained from the regression models are summarized in the middle column of Table 1. As expected, the entries in this column are smaller than those in the other columns because the use of the regression model has effectively removed the contribution to variance due to linear trend in the estimates.

Table 1. Comparison of standard errors using alternative methods

Characteristic of household	Residual from regression	Standard error of model predictions	Using GVF "a" and "b" parameters
Total households	265,375	194,385	258,375
HH with 9-18	201,806	137,821	154,271
HH with 9-18 Northeast	83,955	51,496	70,050
HH with 9-18 Midwest	95,255	59,774	78,103
HH with 9-18 South	106,681	68,143	95,820
HH with 9-18 West	107,700	60,777	77,792

7. Future Research

CPS estimates of the interior cells of the table defined by the raking dimensions are not independent However, we were unable to estimate their covariances from public data sources. We will be exploring with the Census Bureau, ways of estimating the covariance matrix so that this dependence structure can be mimicked in the replicated control totals. Since the cells are likely to have small negative covariances, we think that this refinement will result in smaller estimated NSPY variances, but we do not expect the impact to be large.

8. References

Alexander, C. H. (1987). *A Class of Methods for Using Person Controls in Household Weighting.* Survey Methodology, December 1987, vol. 13, No.2, pp 183-198, Statistics Canada.

Alexander, C. H. (1990). *Incorporating Person Estimates Into Household Weighting Using Various Models for Coverage.* Proceedings of the Bureau of Census Annual Research Conference, pp. 445-662.

Alexander, C. H., and Roebuck, M.J.(1986). *Comparison of alternative methods for household estimation.* Proceedings of the American Statistical Association, Section on Survey Research Methods.

Gunlicks, C.A., et.al. (1987). *Current Population Survey Variance Properties.* Proceedings of American Statistical Association, 1987, Section on Survey Methodology.

Hornik, R., Maklan, D., Cadell, D., Barmada, C., Jacobsohn, L., Henderson, V., Romantan, A., Niederdeppe, J., Orwin, R., Sridharan, S., Chu, A., Morin, C., Taylor, K., Steele, D. (2003). *Evaluation of the National Youth Anti-Drug Media Campaign: 2003 Report of Findings.* Report prepared for the National Institute on Drug Abuse, Washington DC: Westat.

Judkins, D.R. (1991). *National Survey of Family Growth: Design, Estimation, and Inference.* Vital and Health Statistics, September, 1991, Series 2, Data Evaluation and Methods Research; No. 109.

Rizzo, L., and Judkins, D.R. (2004). *Replicate Variance Estimation for the National Survey of Parents and Youth.* Proceedings of American Statistical Association, 2004, Section on Survey Methodology.

Weidman, L., and Bobbit, L., (1991). *How Does Smoothing Estimated Monthly Control Totals Affect SIPP Estimation?* Proceedings of American Statistical Association, (1991), pp. 559-563.