DESIGN AND ESTIMATION FOR LARGE FEDERAL SURVEYS USING ADMINISTRATIVE RECORDS

Fritz Scheuren, Internal Revenue Service

The main purpose of this paper is to discuss how administrative and survey data can be jointly used in making inferences about U.S. income and employment. Despite the paper's title, primary attention will be given to the Current Population Survey (CPS) conducted by the U.S. Bureau of the Census. Proposed changes in this vehicle are examined based on the incorporation of administrative information from the Social Security Administration (SSA) and the Internal Revenue Service (IRS). Most of the comments made in the paper grow out of record linkage research conducted by the Census Bureau with SSA and IRS participation [1].

Organizationally, the paper is divided into three main sections. The first section of the paper contains a few observations on the CPS and some background on the systemic issues in integrating survey and administrative records. Readers should be warned, however, that the intent in this first section is not to give a rounded treatment of the much-studied CPS, but rather to focus on certain aspects of that survey which might be improved by more use of administrative record data. In the second section of the paper, some CPS design and estimation opportunities are suggested that may have some generality and, hence, may be worth considering for other Federal surveys (particularly, the proposed new Survey of Income Program Participation [2]). The final section of the paper is devoted to setting forth a partial research agenda for testing the proposals made.

1. BACKGROUND

In the United States, the history of large-scale government household surveys to measure income and employment has been dominated by the monthly Current Population Survey conducted by the U.S. Bureau of the Census. Even today, the CPS retains many of the features it had when it was started in the 1940's. For example, the number of CPS primary sampling units has greatly increased since the early days; however, the survey continues to have a multi-stage stratified cluster design, relying basically on information in the previous decennial census for its sampling frame. The clustered nature of the CPS poses only minor problems at the National level; for estimates by State, on the other hand, important tradeoffs exist between the efficiency of the estimates and the cost of conducting the survey [3]. The basic structure of the CPS estimation has continued essentially unchanged since the mid- 1950's. Then, as now, a series of ratio adjustments were made. A key part of the approach is to "ratio adjust" weighted survey totals to independent population estimates by age, race and sex. These independent estimates are derived from the previous census by demographic methods [4]. In recent years, questions have been raised about these CPS ratio adjustment techniques because of concerns about the census undercount and the inability to adequately measure net population flows due to international migration [e.g., 5].

Ironically, the CPS may be a victim of its great success. As the premier U.S. household survey, proposed changes have often been viewed with concern (such as more use of the telephone) or have been judged desirable but unnecessary (such as computerizing the household roster to better control the sample). Nonetheless, the Census Bureau is currently making a substantial effort to rethink its approach to the CPS; indeed, many of the ideas in this paper were originally sketched at a Census Bureau-sponsored conference on survey redesign issues which was held last fall [6].

Integrating Administrative and Survey Systems

The relationship between surveys and administrative records can be divided into several phases. These phases reflect a clear, if slowly evolving, pattern toward greater integration which has held for the Current Population Survey, at least, and may hold for other surveys in broad outline.

1. Independent Phase -- The initial phase may be one where the goals of the survey or statistical system and the goals of related administrative system are seen by some, or all, of the parties involved as being unrelated or even opposed. Obviously, the primary focus of an administrative system is to make and record individual program determinations. Statistical summaries are usually required, but, in any event, the production of information is not a major objective of the administrative system [7]. The survey system has, on the other hand, the primary objective of providing information and, in the case at least of Census Bureau surveys, stringent statutory barriers exist against any use of individually identifiable data for program determinations.

2. Complementary Phase -- Despite differences in goals, survey designers use administrative records for sampling frames, to check on survey responses or to augment survey data with information not obtained in the survey but available in an administrative record. In the CPS, for example, use has long been made of local administrative lists of new construction permits to update the survey sampling frame from the

previous census [8]. In the last two decades there have been a number of ad hoc efforts to match CPS data to the administrative records of IRS and SSA [9] so that studies of survey content and coverage could be mounted.

3. Integration Phase -- Full integration of survey and administrative systems is seldom achieved. Differences in objectives generally create statutory, procedural and attitudinal barriers to the attainment of an optimal degree of cooperation. The CPS system is no exception to this general rule, but much more could be done in the way of integration than is presently the case -- with consequent benefits in lowered costs and estimates with smaller mean square errors.

The remainder of this section develops some of the changes in infrastructure on the survey and administrative sides that would be needed to achieve a greater degree of integration.

Survey Infrastructure Changes

The primary change in the CPS would be in the emphasis given to individual and address identification. A completely automated data entry and access system would be needed, linking both the current wave of interviews with those obtained in all prior waves at the same address, e.g., as described in [10]. Identifying information about each household member should be as comprehensive as possible, to maximize the ease of matching addresses and individuals to IRS and SSA records.

Address variations would have to be sought (for instance, 12th and E Streets plus 1201 E Street plus a mailing address, 1111 Constitution Avenue, all represent the same address [11]). Most important for addresses would be the ZIP code designation. Software already exists that allows for ZIP code assignments when missing [12]. Use of such software would be one ingredient in the address perfection process. It is also necessary to explore the possibility of asking for an address as of a particular point in time (say as of April 15th) or asking for the last address prior to the present one. Obviously, 9-digit ZIP codes could be immensely valuable if they begin to be used extensively. Use of telephone numbers, as identifiers, may also be valuable, since these may soon be required on more and more administrative records.

Full names and social security numbers (SSN's) would be an obvious requirement for personal identification. Sex, race, date of birth, place of birth, maiden names, and nicknames should all be sought, as well. While this list may seem formidable, experience with the Income Survey Development Program suggests that over 90% of the SSN's of adults could be obtained fairly readily [13]. (Right now, with very little emphasis, the CPS is obtaining usable SSN's for between 70% and 80% of all interviewed adults.)

A system of matching the survey addresses and individuals to the administrative record files would have to be developed. Since such a system might well have to be carried out under special custody and access arrangements, it is described separately below.

Administrative Infrastructure Changes

Three kinds of changes would be needed. First, the regulations now governing Census Bureau access would have to be altered, to expand the IRS and SSA information made available to the Census Bureau for the sampled individuals and to enlarge the uses to be made of this information [14]. Access to information returns on unemployment compensation (Forms 1099UC) filed by the States would be an obvious need. Second, certain information now obtained on the tax return (e.g., occupation [15]) but not used electronically, might be captured -- at least for a large sample of taxpayers, including, of course, those in the CPS. Residence information now obtained irregularly might also need to be provided with greater frequency (albeit perhaps only for taxpayers who move or whose mailing and residential addresses are different). Third, if the CPS becomes strongly dependent on SSA and IRS data, then a timely reliable delivery system with guaranteed quick turnaround will have to be developed. In a monthly survey, delay is disaster.

Matching Infrastructure Changes

At the present time, both IRS and SSA have their own computerized systems for finding missing and misreported social security numbers. IRS relies heavily on surname and current address; SSA, on surname and certain demographic information, like date and place of birth. Both approaches appear adequate for administrative uses but may not be entirely appropriate for statistical work. Problems of multiple matches arise which can involve costly clerical intervention. Also, there is no internal method of determining the extent of erroneous nonmatching that might arise because of response errors.

Ample guidance is available in the statistical literature [16] on up-to-date matching algorithms that, with modifications, could be applied to the IRS and SSA systems. Such algorithms could automate the selection among multiple possible matches, although some clerical review would still be advisable. To deal with the problem of estimating erroneous nonmatches, a capture-recapture [17] or multiple systems approach [18] could be taken, where an attempt is made to independently match, using each of the existing IRS and SSA systems (and perhaps others, as well).

2. OPPORTUNITIES FOR IMPROVED CPS DESIGN AND ESTIMATION

This section sketches out the increased opportunities for improved survey design and estimation obtainable through the use of administrative records. Two main options can be distinguished, depending on whether or not exact

matching is carried out to the administrative data.

Potential Benefits in the Absence of Exact Matching

Two areas that can benefit from the use of administrative totals are geographic post-stratification of survey estimates for variance reduction and small area estimation. The geographic post-stratification being referring to is similar to the first-stage ratios now used in the CPS but could be done with more up-to-date and potentially more highly correlated administrative data, leading to definite improvements in both income and employment estimates at the State level. (Improvements at the national level would undoubtedly be minor though.)

For example, administrative totals from the Unemployment Insurance (UI) system [19], might be used at the county level in the design of the CPS, where the sample design would be modified using administrative records for formation of PSU's, stratification, and sampling measures [20]. Administrative variables that are well correlated with the items to be measured could be made available by county for the formation of PSU's, and strata established for sampling. Use of administrative data would also permit better estimates of the PSU component of the variance, something that might be of special value in the smaller States [21].

As the variables to be used would be available on a regular, perhaps annual, basis, it may be possible to redesign or resample for large Federal sample surveys more often than once every decade, leading to more efficient surveys in the Federal government, with a lower cost for the surveys.

Possible Benefits With Exact Matching

The above examples all involve administrative totals that can be made readily available without matching between the record system and the survey. With matching, other gains can be realized. The first and most obvious benefits would be from use of ratio estimators to reduce sample variances for State or regional estimates. The UI system mentioned above could be used in the CPS for this type of application, since unemployment compensation payment information on each recipient is reported to IRS annually (on the Form 1099UC); health surveys and other major surveys conducted by the Federal government could also benefit from the use of more highly correlated variables for ratio or regression estimation.

The current (second-stage) CPS post-stratification by age, race and sex could also be improved by the combined use of SSA and IRS information. Experimental work on this latter application has already been completed successfully [22], and the idea offers much promise, especially if combined with the direct use of UI data mentioned above (possibly after matching the UI information to Social Security records to obtain age, race and sex).

An important design modification available through the use of administrative records would be the increased use of multiframe designs. Multiframe designs offer three advantages over traditional sample designs: improved coverage of groups not represented adequately in area sampling frames, the ability to oversample groups of special interest that may be rare in the population, and reduced sample variance on the estimates generated from the survey. Multiframe designs using IRS and SSA records as a starting point are particularly attractive for the Survey of Income and Program Participation, where some experimentation has already taken place [23].

A third use of matched data would be methodological; administrative data could be used in an ongoing analysis of the reliability of survey data. In pretesting of survey methods, administrative records have been used in victimization and health surveys to determine the completeness of reporting for various methods of data collection. The Bureau of the Census has used records from police stations to study the completeness of reporting of crimes, records from utility companies to determine completeness of reporting of energy usage, and tax records and program records from the Social Security Administration (SSA) to study reports of earnings. These studies focused not only on the completeness of the reporting on the surveys, but also on the validity of the reports. Models of misclassification or underreporting could be developed to provide survey estimates subject to less bias due to response and nonresponse errors [e.g., 24].

3. CONCLUDING COMMENTS

This paper has touched briefly on changes in Federal survey-taking that would permit greater use of administrative records, especially those from SSA and IRS. At various points conjectures are made about the efficiencies that more use of such records might create. Evidence exists to suggest that the cost of adopting at least some of the recommendations made here could well be worthwhile. The real challenge, however, is to choose from among the opportunities and options those to do more research on in the near future.

A joint research effort is needed, led by Census Bureau staff, but with strong participation from statisticians at the Department of Labor, Social Security, and the Internal Revenue Service. It is recommended that the following three areas be among the first studied:

1. Greatly increased emphasis should be placed on obtaining identifying information in the CPS. Internal survey management procedures should be upgraded, as in the Canadian Labour Force survey [10], so that the identifiers are readily available for matching to administrative records. The panel structure of the CPS sample should also be used to perfect this identifying information when it proves incomplete or inaccurate.

2. Social Security numbers (SSN's) should be obtained routinely in the CPS, not just in

the March survey. This should be done even though the information would not be immediately used in the survey estimation process. For one thing, by collecting good SSN's and related identifiers, ongoing efforts to produce a data base for epidemiologic research [25] could be greatly advanced; furthermore, when a new estimation procedure was perfected, it could be applied retrospectively, to track how the new and old estimation procedures behave over time.

3. Off-line experiments should begin on integrating IRS and SSA administrative data directly into the current March CPS procedures. (Work on integrating such approaches into the Survey of Income and Program Participation might also be carried out, but this is not recommended now because of the normal start-up problems that survey is likely to encounter.)

   The single most important administrative record system to employ in post-stratifying the CPS would be a combination of the wage (W-2) and unemployment compensation (1099UC) data bases matched to Social Security age, race and sex information. CPS State data on both employment and unemployment should show marked reductions in bias and variance with such a strategy. (Use of these same record systems to upgrade unit and item nonresponse adjustments would be an important second priority.)

Any research on improving Federal surveys must also take into account parallel developments involving changes in the use of administrative records for small area intercensal estimation and possible greater reliance on administrative records in the decennial census itself. Perhaps on another occasion we can treat these related research activities as part of an integrated whole.

ACKNOWLEDGEMENTS AND AFTERWORDS

Thanks are due to a number of people who helped in the preparation of this paper: in particular, to Charles Cowan at the U.S. Bureau of the Census, who provided a number of the ideas in section 2. Others who helped at the Census Bureau included Gary Shapiro and Charlie Jones. At IRS thanks are due to H. Lock Oh, for his editorial assistance, and to Renee Wheeler and Denise Herbert, for their typing support.

The views expressed in this paper do not reflect the official position of the U.S. Bureau of the Census or of the Internal Revenue Service. They are simply a small addition to the continuing dialogue on how to increase the uses of administrative records for statistical purposes. Any ambiguities or errors in this paper are entirely the responsibility of the author.

NOTES AND REFERENCES

[1] Kilss, Beth, and Fritz J. Scheuren, The 1973 CPS-IRS-SSA Exact Match Study, Social Security Bulletin, Vol. 41, No. 10, 1978.

[2] Herriot, Roger A., The Use of Administrative Records in Social and Demographic Statistics, paper presented at the biennial meeting of the International Statistical Institute, Madrid, September 1983.

[3] In some of the smaller States, the nonself-representing PSU's make up an important part the estimate; also, there are only a small number, leading to the possibility that the clustering will heavily impact on the variance.

[4] Hanson, Robert H., The Current Population Survey: Design and Methodology, U.S. Bureau of the Census, Technical Paper 40, 1978.

[5] Lancaster, Clarise, and Fritz J. Scheuren, Counting the Uncountable Illegals: Some Initial Statistical Speculations Employing Capture - Recapture Techniques, 1977 Proceedings, American Statistical Association, Social Statistics Section, pp. 530-535.
Warren, R., and J. S. Passel, Estimates of Illegal Aliens from Mexico Counted in the 1980 Census, presented at the annual meeting of the Population Association of America, Pittsburgh, April 1983.

[6] U.S. Bureau of the Census, Redesign Advisory Panel Meetings, October 1982.

[7] The modern income tax law did not provide that statistical summaries be prepared until 1916, 3 years after its inception in 1913. The social security system, on the other hand, seems to have given early and, until recently, substantial support to the production of statistical information from its files.

[8] Hanson, Ibid.

[9] U.S. Dept. of Health and Human Services, Series on Studies from Interagency Data Linkages, 1973-80, Social Security Administration.

[10] Ashraf, Anis, and Ian Macredie, Edit and Imputation in the Labor Force Survey, 1978 Proceedings, American Statistical Association, Survey Research Methods Section, pp. 425-430. Depending on the options discussed in section 2, some of these infrastructure changes may not be needed.

[11] These addresses describe the location of the Statistics of Income Division at IRS.

[12] The Internal Revenue Service makes extensive use of such software in "perfecting" ZIP codes, since postal rates are lower on presorted material and IRS is among the largest of the mass mailers.

[13] Kasprzyk, Daniel, Social Security Number Reporting, The Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program, Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program, pp. 123-144, Social Research Council, New York, 1983.

[14] Alvey, Wendy, and Fritz J. Scheuren, Background for an Administrative Record Census, 1982 Proceedings, American Statistical Association, Social Statistics Section.

[15] Crabbe, Patricia, Peter Sailer, and Beth

Kilss, Occupation Data from Tax Returns: A Progress Report, 1983 Proceedings, American Statistical Association, Survey Research Methods Section.

[16] Smith, Martha E., Development of a National Record Linkage Program in Canada, 1982 Proceedings, American Statistical Association, Survey Research Methods Section, pp. 303-308.

[17] Bishop, Y. M., S.E. Fienberg, and P. W. Holland, Discrete Multivariate Analysis: Theory and Practice, pp. 229-256, MIT Press, Cambridge, 1975.

[18] Marks, E.S., W. Seltzer, and K. J. Krotki, Population Growth Estimation, A Handbook of Vital Statistics Measurement, The Population Council, New York, 1974.

[19] U.S. Dept. of Commerce, Statistical Policy Working Paper 6: Report on Statistical Uses of Administrative Records, Office of Federal Statistical Policy, 1980.

[20] For an excellent summary of similar work being conducted in Canada using the Canadian UI System, see Leyes, John, Ellen Bobet, and Louise Radley, The Use of Unemployment Insurance Records to Derive an Unemployment Indicator, 1982 Proceedings, American Statistical Association, Survey Research Methods Section, pp. 284-287.

[21] The estimation of variances at the State level is very difficult in the CPS because of the small number of PSU's involved. Models relating exact variances for the first stage of the design obtained from administrative data to direct estimates from the CPS might yield more stable results than current methods. It might be added that variances could also be reduced, as well as better estimated. Modeling discrepancies between related (but not identical) variables in the administrative and survey environments may aid in detecting bad interviewing and help focus reinterview resources and possible remedial actions.

[22] Scheuren, Fritz, Chapter 1, Studies from Interagency Data Linkages, Report No. 10, Social Security Administration, 1981.

[23] Ycas, Martynas A., and Charles A. Lininger, The Income Survey Development Program: Design Features and Initial Findings, Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program, pp. 25-32, Social Research Council, New York, NY, 1983.

[24] Little, Roderick J. A., and Michael E. Samuhel, Alternative Models for CPS Income Imputation, 1983 Proceedings, American Statistical Association, Survey Research Methods Section.

[25] Rogot, E., M. Feinleib, K.A. Ockay, S.E. Schwartz, R. Bilgrad, and J.E. Patterson, On the Feasibility of Linking Census Samples to the National Death Index for Epidemiologic Studies, American Journal of Public Health (in press).