Volume Title: Annals of Economic and Social Measurement, Volume 2, number 2

Volume Author/Editor: NBER

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/aesm73-2

Publication Date: April 1973

Chapter Title: The Creation of Longitudinal Data from Cross-Section Surveys: An Illustration from the Current Population Survey

Chapter Author: Terence F. Kelly

Chapter URL: http://www.nber.org/chapters/c9892

Chapter pages in book: (p. 209 - 214)

Annals of Economic and Social Measurement, 2/2, 1973

THE CREATION OF LONGITUDINAL DATA FROM CROSS-SECTION SURVEYS: AN ILLUSTRATION FROM THE CURRENT POPULATION SURVEY

BY TERENCE F. KELLY*

INTRODUCTION

That the Current Population Survey (CPS) is a unique data source is relatively well known. The CPS provides nationally representative data on demographic, social, and economic circumstances of U.S. families and individuals thereby supporting rather elaborate cross-section analyses of one sort or another. Since available tapes go back to 1960, they now form the basis for time series analyses as well. And, of course, the combination of these two factors allows pooling of cross-section and time series observations thereby furnishing the analyst with an extremely powerful body of micro-data.[1] Finally, used in conjunction with decennial censuses, the CPS can be used to conduct extremely detailed cohort analyses.

A lesser known feature of the CPS is the fact that it can be used to generate a modified longitudinal micro-data set.[2] Some three-quarters of the CPS is common from month to month and 50 percent from year to year, meaning that it is possible to generate a subsample containing the same set of households across successive months or two-year periods. Such longitudinal data are useful to both analysts and to policymakers who might be interested in generating dynamic estimates of response patterns, as opposed to static cross-section or short-run time series estimates.

There are, of course, other sources for longitudinal data.[3] One advantage of the CPS, however, is that it provides a data set covering the entire period since 1959,

* Data upon which this study was based were generated with funds from the Office of Economic Opportunity under Contract BIC-5244. All data processing was performed by the Hendrickson Corporation. A lengthier, more detailed version of this paper may be obtained upon request to the author.

[1] The term "micro-data" is used to refer to observations on individual units, rather than on groups of units. Further discussion of micro-data may be found in Malcolm S. Cohen, "The Micro Approach to Manpower Research," Industrial Relations Research Association, *Proceedings of the Twenty-First Annual Winter Meeting*, Chicago (December 1968), pp. 120–128.

[2] The term "longitudinal" is used to indicate a sample containing the same points of observation over time. The CPS is a modified longitudinal sample because the period containing the same households is relatively short and because, as will be discussed below, in certain instances it is not entirely certain that the data set actually contains the same families. Previous studies utilizing the longitudinal component of the CPS include Glen G. Cain and James D. Smith, "Markov Chain Applications to Household Income Distribution." Paper presented at the 1967 meetings of the Econometric Society, Washington, D.C.; Terence F. Kelly, "Factors Affecting Poverty: A Gross Flow Analysis," President's Commission on Income Maintenance Programs, *Technical Studies*, Washington, D.C., U.S. Govt. Printing Office, 1970, pp. 1–81; and Marshall L. Turner, Jr., "A New Technique for Measuring Household Changes," *Demography*, Vol. 4, No. 1, 1967, pp. 341–350.

[3] Three recent longitudinal surveys are the Office of Economic Opportunity's Survey of Economic Opportunity, the Ohio State Longitudinal Survey, and the University of Michigan, Institute of Social Research, Panel Study of Income Dynamics. One of the better known longitudinal data sources is the Social Security Administration's Continuous Work History File. See Lowell Gallaway, "The Negro and Poverty," *The Journal of Business*, Vol. 40, No. 1, January 1967, 27–35.

longer than many other sources. In addition, the sample includes the entire U.S. population, rather than being limited to particular demographic subgroups. On the other hand, the CPS matched data suffer from certain limitations, as will be noted subsequently.

Matching the Current Population Survey Over Time

There are a variety of matches possible with the CPS. One may match the responses of persons or of families, the match may cut across months or across years, and it may be restricted to a particular recurrent supplement to the normal CPS. This paper is concerned with matching responses of families in the February–March Supplement from year to year over the period 1960 to 1969.[4] The CPS is carried out each month. The monthly surveys consist of two components. The first is a recurrent portion which provides data on the monthly labor force and employment status of the non-institutionalized U.S. population. These data are published by the Bureau of Labor Statistics of the Department of Labor, providing the rather familiar monthly estimates of unemployment and labor force rates and levels. The second part of the monthly CPS is known as the supplement series. Contents of the supplements vary from month to month and occasionally the supplements are devoted to special, one-shot topics. The October supplement, for example, is devoted to the question of employment and unemployment of school age youth. In May, the supplement covers multiple job holding.

The February supplement is devoted to questions about the work experience of the population during the year prior to the survey.[5] This yields information on such topics as weeks worked, weeks of unemployment, amount of part-time work, reasons for non-participation in the labor force and so on. Then in March, the supplement covers questions on income in the year prior to the survey. These data underlie the familiar income and poverty status tabulations published by the Census Bureau. Thus, the combined February–March supplement provides data on various aspects of work and income along with such demographic information as race, age, sex, education and household location.

In order to create a continuous matched data set over time it was first necessary to reformat some of the initial February–March CPS files. In 1967 the Census Bureau altered and expanded the CPS. As a result the format of the 1968 and 1969 tapes is quite different from the ones used previously. In order to perform matches and to have consistent output tapes, new 1968 and 1969 tapes were created in the format used in earlier years. These reformatted tapes were then used as input to the matching program. Further, the expanded survey caused some changes in the fields used to identify matching families, so a slightly different version of the 1967 tape is

[4] 1960 is the earliest CPS tape available, and 1969 is the latest tape I had at my disposal. More recent tapes will be added to the matched set as they become available. Analysis was restricted to family records due to cost considerations. The Social Security Administration and the Office of Economic Opportunity are sponsoring a project which will involve matching of person-family records. These matched data will provide information on families and on the persons within those families.

[5] After 1971, work experience questions were asked in April for a subcomponent of the survey.

necessary for the 1967–1968 match. In short, three sets of formats were merged into one.[6]

Part of the CPS sample is changed each month in order to avoid problems of lack of cooperation which might arise when a constant panel is interviewed indefinitely. A household is in the sample for four consecutive months one year, leaves the sample during the following eight months, and returns for the same four calendar months next year. Under this system, some 75 percent of the sample is common from month to month and 50 percent from year to year. Since each household is identified by a special code, it is theoretically possible to "match" the responses of three-quarters of the sample from month to month or one-half from year to year.[7] In creating a matched set of February–March files, then, there are two matches to be performed: the first across the successive two months and the second across successive years. Of the two, the former is the less troublesome. Prior to 1967, the Census Bureau performed the February–March match prior to releasing the tapes, indicating those households which could not be matched. Since 1967, work experience information is collected for all households in the March supplement. Those families entering the sample for the first time during March are asked work experience information in April. Therefore, there is actually no need to match between February and March since it has already been done by the Census Bureau.

Matching across years is a considerably more intricate operation, particularly in the case of the family files. It has been mentioned that the files contain scrambled identifiers which, on the face of it, should make the process relatively simple. Unfortunately, however, a family unit is given the same identifier so long as even one of its members is common across the sample period. This means that a given eight-person family could lose seven members in one year who are then replaced by seven new members during the following year, and yet the family would still be given the same identifying number! In order to screen peculiarities such as these from the files, a rather elaborate set of decision rules was developed. These rules are presented in flow chart form in Table 1. As may be seen, the rules essentially involve checking on a variety of demographic characteristics: family type, race, age, sex, and marital status.

The matched sample is considerably reduced in size by the addition of these household composition checks. It is further depleted by the fact that many families are not interviewed in successive March periods. Many families move, are not at home when the interviewer calls, or simply refuse to respond to the questionnaire. The net effect of these factors is to reduce the matched sample to about one-third of the size of the full CPS in most years.[8] Table 2 presents a count of families and unrelated individuals in the matched and full CPS March sample from 1960 to 1969.

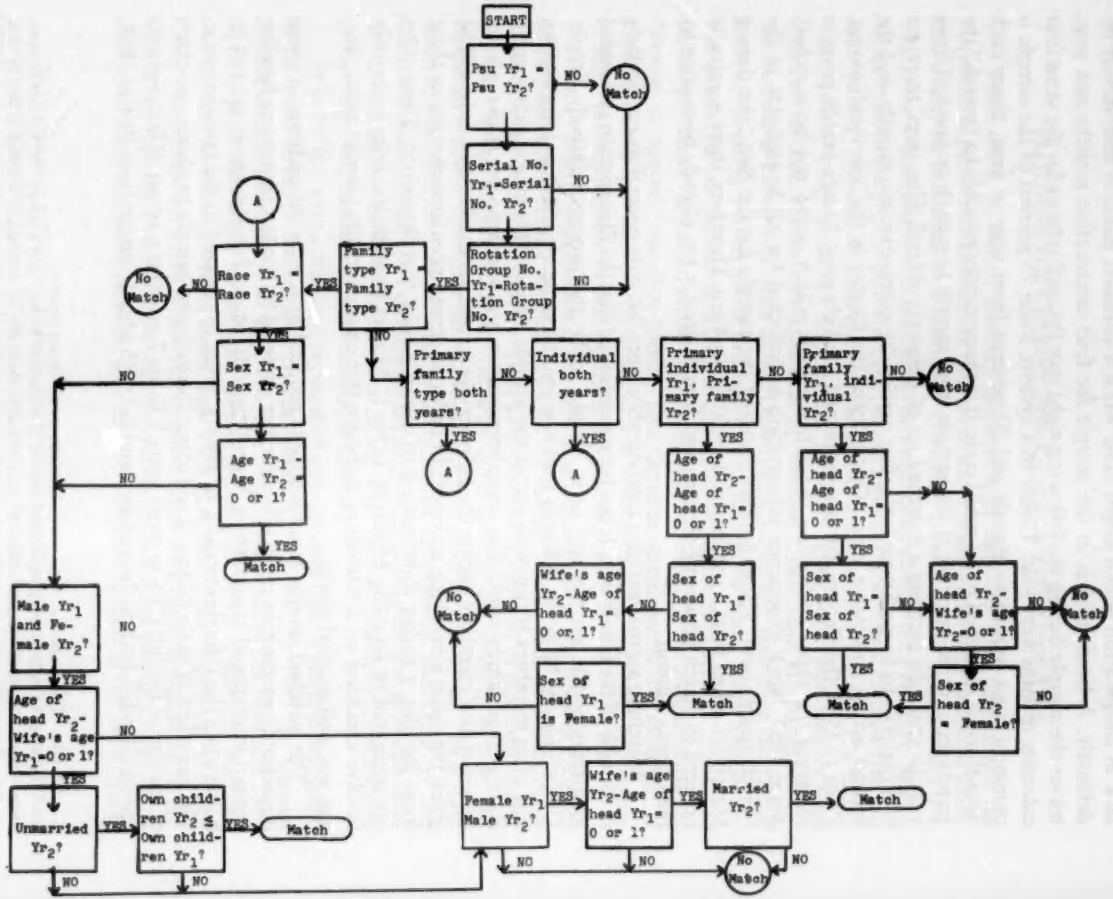
[6] Copies of the common format will be made available on request.

[7] In order to protect the confidential nature of responses by any given family, the Census Bureau "scrambles" the identification codes. It thus becomes possible to determine whether a family is a "match" without being able to identify confidential family characteristics, such as name or address. Unfortunately, matches can only be made over successive two-year periods—it is not possible to follow families for more than two years.

[8] This is not true in the 1961–1962 match which is an anomalous case to be discussed subsequently.

TABLE 1

DECISION RULES USED TO MATCH FAMILY FILES



Evaluation of the Matched Data[9]

In order to evaluate the quality of the matched data, a number of validity checks were made. Tests of accuracy and validity were undertaken, using the overall March CPS as a standard of comparison. Additional comparisons were made between the matched data file and the responses of families whose records, for one reason or another, could not be matched.[10] Finally, the matched file was subjected to a variety of special edits.

In general, the matched data appeared quite representative of the full CPS, at least in terms of demographic composition. Further, there were no significant differences in the demographic structures of the matched and unmatched data. In other words, there is no evidence of significant bias in the matched data. Two particular problems should be noted, however.

First, as is apparent from Table 2, the 1962–1963 match is considerably reduced in size. Every decade, the CPS is benchmarked to the decennial census of the population to account for geographic shifts of the population. At the time of benchmarking, certain new sample points are added and others are dropped. This accounts for the inability to match responses between 1962 and 1963. Those responses which could be matched were compared with the full CPS and not found to be significantly different along most dimensions, but the reduced sample size may be insufficient to support certain detailed analyses.

TABLE 2

Size of Matched February–March CPS Files, 1960–1969

Match Years	Full CPS (households)		Failed Due to Composition Change		Failed Due to Income Edit	Number of Final Matches	
	Number	%	Number	%		Number	%
1960–61	34,620	100	8,979	25.9	1,407	8,460	24.4
1961–62	33,350	100	8,774	26.3	658	7,732	23.1
1962–63	32,924	100	11,622	35.3	0	2,028	6.1
1963–64	24,290	100	8,334	34.3	0	8,268	34.0
1964–65	24,284	100	8,935	36.8	0	8,619	35.4
1965–66	24,490	100	9,579	39.1	0	8,549	34.9
1966–67	33,500	100	14,934	44.6	0	10,242	30.5
1967–68	46,807	100	22,851	48.8	0	17,197	36.7
1968–69	43,362	100	22,530	46.6	0	17,790	41.0

Second, it should be stressed that the matched subfile is really a sample of the *nonmobile* population. As such, it can be expected to differ slightly from more general samples. For example, a comparison of the annual percentage change in

[9] Potential limitations of CPS matched data have been discussed by Harvey J. Hilaski, "The Status of Research on Gross Changes in the Labor Force," U.S. Department of Labor, *Employment and Earnings*, Vol. 15, No. 4, October 1968, pp. 6–13; Susan Palmer, "On the Character and Influence of Nonresponse in the Current Population Survey," American Statistical Association, *Proceedings of the Social Statistics Section 1967*, pp. 73–80; and Robert B. Pearl, "Gross Changes in the Labor Force: A Problem in Statistical Measurement," U.S. Department of Labor, *Employment and Earnings*, April 1963, pp. ix–xx.

[10] The chi square test was used to determine the correspondence of alternative percentage distributions. Such distributions as age, sex and race of head, and region were compared.

poverty among families and unrelated individuals using the matched data set with the full CPS shows some pronounced divergencies. In certain years, namely those between 1963 and 1966, the matched data show an increase in poverty where the overall CPS shows a decline. A likely explanation of this phenomenon is that the matched sample fails to represent adequately those families whose incomes increase substantially, meaning that it biases downward estimates of income improvements. Families with large income increases are probably more likely to move than are other families and of course, if they move they cannot be included in the matched sample. We might also conjecture that the converse holds: families with large decreases in income are probably more likely to move than are families with stable incomes. On net, these considerations suggest that the estimated variance in year to year income changes is smaller in the matched data than one would expect for the total population.

Additional edit checks were performed on the tapes. Certain years contain missing data or wild income codes.[11] As long as the potential user is aware of these difficulties, there should be no problem in using the files.

Conclusion

Copies of the matched data tapes can be obtained from the author. The procedure to be followed is the same as that outlined by Jodie Allen in this issue. The data are unique, since they provide a basis for modified longitudinal analyses of the U.S. population. There are, quite naturally, limitations with the data set. Some appear to be relatively serious; others are minor. The first drawback is the fact that matches can cover only two years. Two years is a rather small slice out of the life of a family and for certain analyses it may be too short. Other limitations which have been investigated include the reduced size of the 1962–1963 match and potential biases. In comparing the matched data with the full CPS and with unmatched data, certain discrepancies were uncovered but it would appear that the problems are tractable. If the analyst is willing to limit comparisons to relatives (e.g. males compared to females or whites compared to nonwhites), the CPS matched data file can be a powerful analytical tool.

The Urban Institute

[11] See Jodie T. Allen, "A Guide to the 1960–1971 Current Population Survey Files," this issue.