LONGITUDINAL LABOR MARKET DATA: SOURCES, USES, AND LIMITATIONS*

Orley Ashenfelter and Gary Solon
Princeton University

*This paper is a revision of a paper presented at the conference on "An Assessment of Labor Force Measurements for Policy Formulation," sponsored by the National Council on Employment Policy, Washington, D.C., June 17, 1982. The authors thank Burt Barnow, Michael Borus, Sar Levitan, Michael Ransom, and the conference participants for comments and information.
LONGITUDINAL LABOR MARKET DATA: SOURCES

USES, AND LIMITATIONS

Until recently, most research on labor force behavior and experience analyzed cross-section data, which pertain to a population sample at a single point in time. Recent years, however, have seen the development of several longitudinal data bases, which follow the same individuals over multiple points in time. Two factors have contributed to the development of longitudinal information. One is that convincing research on a number of public policy issues requires longitudinal data. Indeed, without longitudinal data, some important research issues cannot be addressed at all. For example, appropriate public policy towards poverty, unemployment, and welfare dependence depends partly on whether families' or individuals' experience of these states is typically transitory or chronic. Cross-sectional snapshots of the poor or the unemployed, which focus on different individuals at different times, cannot possibly reveal how many of those poor or unemployed at one time remain poor or unemployed at later dates. Such questions of state persistence necessitate longitudinal tracking of the same individuals. Still other issues previously addressed with cross-section data can be treated with more reliable research methods when longitudinal information is available. For example, longitudinal data have enabled more thorough investigations of the effects of government training programs on earnings and the wage impact of union status.
The second factor is that the cost of developing useful longitudinal data sets is no longer prohibitive. In some cases, computerized matching of existing administrative records can produce inexpensive longitudinal information. In other cases, valuable longitudinal data bases can be generated by computerized matching of existing administrative and survey data. Even where the desired longitudinal information can be collected only by initiating new surveys, the advance of computerized data management systems has made longitudinal data development cost-effective in the last 15 years.

The purpose of this paper is first to describe briefly the major sources of longitudinal data and their relative merits. The discussion then turns to a review of the types of analysis for which longitudinal information has proven especially useful.

Sources of Longitudinal Labor Market Data

Longitudinal labor market data have been generated in three main ways. The first is longitudinal matching of administrative records on participants in government programs. The most prominent example is the Social Security Administration's Continuous Work History Sample (CWHS). This data set contains longitudinal earnings records for a sample of workers covered by the social security program. Another example is the Labor Department's Continuous Wage and Benefit History (CWBH), which contains longitudinal information on the earnings, benefit experience, and
other characteristics of a sample of workers covered by unemployment
insurance. The Labor Department has also assembled the Continuous Longitu-
dinal Manpower Survey (CLMS) data, which consist partly of administrative
information on a sample of enrollees in programs funded under the Compre-
hensive Employment and Training Act.

Surveys are a second source of longitudinal data. Longitudinal sur-
evay data can be collected either through one-time retrospective surveys
that obtain information on individuals' past experience or through panel
surveys that periodically reinterview the same individuals. The latter
approach is exemplified by the National Longitudinal Surveys (NLS) of labor
market experience and the Panel Study of Income Dynamics (PSID). The NLS
project, conducted for the Labor Department by the Census Bureau, the
National Opinion Research Center, and Ohio State University's Center for
Human Resource Research, has followed samples of several age-sex cohorts:
men of age 45 to 59 in 1966, men 14 to 24 in 1966, women 30 to 44 in 1967,
women 14 to 24 in 1968, and men and women 14 to 21 in 1979. The original
sample size for each of the 1960s cohorts was about 5,000 individuals, and
the 1979 cohort started with over 12,000 individuals. The wide variety of
information collected by NLS, as Michael Borus [3] put it, "includes
everything you always wanted to know about individuals that the Census
Bureau was not afraid to ask."

The PSID effort, initiated by the Department of Health, Education,
and Welfare and conducted by the University of Michigan's Survey Research
Center, has collected since 1968 a similarly wide variety of information on
a national sample of families that overrepresents low-income families. As some of the original 4,800 families have split and rearranged, PSID has interviewed the originally sampled individuals' new family units so that, despite sample attrition, the sample has actually grown over time. Other special longitudinal surveys, described elsewhere,1 include the Longitudinal Retirement History Study, the National Longitudinal Study of the High School Class of 1972, Project Talent, High School and Beyond, and the NBER-Thorndike-Hagen survey.

Another important panel survey is the Current Population Survey (CPS), the monthly national household survey by the Census Bureau that produces the unemployment rate and other regular labor force statistics. Although the CPS is usually viewed as a source of cross-section and time series data, it has a panel aspect as well. A household whose address is selected for the survey is interviewed for four consecutive months, dropped from the survey for eight months, and then interviewed for another four months before leaving the sample for good. It is therefore possible to match the survey responses of a household for up to a 16-month period (unless the household moves from the selected address, in which case the household that moves in is interviewed in its place). Compared to the NLS and PSID data, the CPS longitudinal information spans a shorter period, contains fewer variables, and does not follow movers, but it pertains to a much larger sample and, unlike NLS, represents all demographic groups.

A third source of longitudinal labor market data is the series of negative income tax experiments conducted since the late 1960s. Each of
these experiments — conducted in New Jersey and Pennsylvania, Seattle and Denver, Iowa and North Carolina, and Gary, Indiana — set up a pilot negative income tax program lasting several years for a selected experimental group, and also observed a control group over the same period. The main purpose was to compare the labor supply behavior of the two groups to estimate the likely work incentive effects of a national negative income tax. The data also can be used more generally to explore patterns of welfare dependence and labor market experience among low-income families.

Before considering the analytic uses of longitudinal data in general, it is worth mentioning a few of the relative advantages and limitations of different sources of longitudinal data. One important comparison is between administrative and survey data. In cases where administrative files contain the desired data on the appropriate population, the advantages of administrative data are considerable. To begin with, longitudinal collation of data already collected in the process of program administration is less expensive than generating the data with new surveys. Consequently, longitudinal data bases from administrative sources often include larger samples than surveys can feasibly interview. Also, during the period of the sample's program participation, administrative data are relatively free of the problems surveys have with nonresponse and sample attrition. In addition, information from administrative records may, in some cases, be more accurate than information elicited from survey respondents. Survey data on income, for example, are sometimes unreliable. A comparison by Herriot and Spiers [7] of CPS and Internal Revenue Service
data on earnings of the same individuals showed discrepancies of at least 15 percent between the two sources of earnings information for almost 30 percent of the matched sample. Despite the likelihood of income underreporting in the IRS data, the CPS earnings data tended to be even lower. Earlier matched comparisons of CPS and census data, initial and reinterview census data, and census and tax data found similar evidence of income measurement error in surveys.\(^2\)

On the other hand, whether administrative files do contain the desired data on an appropriate sample is a big "if". The information collected for administrative purposes is typically narrower than what is desired for research purposes. The CWHS data, for instance, include only a few variables besides earnings, and even earnings are measured only up to the social security taxable limit. The sparseness of administrative information has led the CWBH and CLMS projects to supplement their administrative data with information collected in interviews or questionnaires.

Furthermore, administrative data may not correspond to the population of interest. The CLMS data, for example, are insufficient by themselves for evaluating the impact of training programs on earnings because the data pertain only to program enrollees. A proper evaluation also requires information on a control group not enrolled in training programs. Analysts of the CLMS data have resorted to CPS data matched with social security earnings records to obtain control group information.
In cases where some or all of the desired longitudinal data must be gathered in surveys, it becomes important to consider the merits of retrospective versus panel surveys. Of course, obtaining longitudinal information retrospectively in a single interview is less costly than repeated interviewing. The retrospective, single-interview approach also eliminates sample attrition and yields longitudinal information more quickly. Furthermore, retrospective data are less susceptible to some types of response error. If, for example, a panel survey respondent describes the same job differently in successive interviews or if different interviewers code the same job differently, the respondent may be erroneously recorded as having changed occupations. This sort of error is less likely to occur if the information is collected in a single interview.

On the other hand, panel surveys are less subject to recall error. A retrospective survey respondent that changed jobs five years ago may fail to recall the old job or may forget when the job change occurred. Furthermore, a retrospective survey respondent's recollections might be biased by subsequent events. Of course, just as longitudinal data bases sometimes contain both administrative and survey information, longitudinal surveys can fruitfully combine the retrospective and panel approaches. Indeed, panel surveys typically do collect information retrospectively for periods before and between interviews.

Finally, where a panel survey has been initiated, an important question is how long to continue the survey. This question has arisen recently with regard to whether the 1960s NLS cohorts, originally planned
to be interviewed for 15 years, should be followed for another five years. The answer depends partly on the advantages of having a 20-year, rather than a 15-year, longitudinal history. Another consideration is that continuation of an existing longitudinal survey is a relatively inexpensive way to obtain current data. Even if the new data will be used largely for cross-section analysis, collecting the data from an ongoing panel avoids the costly process of selecting a new sample and developing a new data-processing system.

This advantage is at least partly offset, however, by the sample attrition problem. By 1981, all four of the NLS surveys started in the 1960s had lost at least 1/4 of their original samples. Such attrition not only reduces sample sizes, but, if sample leavers differ systematically from sample stayers, it also might cause the remaining samples to be unrepresentative of the corresponding populations. Even in the PSID project, where sample sizes have grown over time because the survey incorporates new family units containing original sample members, it is unclear how well the current sample represents any population of interest. Therefore, while extending panel surveys generates new data economically, it may do so at a cost of progressively less representative samples. This raises the difficult question of when it is optimal to begin a new survey as opposed to continuing an old one.
Uses of Longitudinal Data

Longitudinal data are particularly advantageous for three types of research: the measurement and analysis of changes in individuals' status over time, the analysis of intertemporal relationships, and analysis that controls for unobserved variables. Although this list of uses may seem abstract, examples of each type will show that these longitudinal analyses often have considerable practical relevance. The examples are intended to serve as illustrations of the kinds of research enabled by longitudinal data, not as an exhaustive compilation of the findings of longitudinal research.

Measurement and Analysis of Change

Cross-section data can tell what proportion of the labor force is unemployed or describe the distribution of wage rates or family income at a point in time. In addition, time series of aggregated cross-section data are useful indicators of general trends and cyclical patterns in unemployment, wages, income, and so forth. Neither cross-section nor time series data, however, can tell how many of those unemployed in one month find employment in the next month or how individuals' wage rates or incomes change over time. Only longitudinal data, which track the same individuals over time, can measure such changes.
An illuminating example is the gross flow data from the CPS. These data show not only how many of one month's unemployed are employed the next month, but also the magnitudes of all the other month-to-month flows among employment, unemployment, and nonparticipation in the labor force. Furthermore, the underlying data on individuals' changes in labor force status can be analyzed to identify the determinants and correlates of transitions among labor force categories. For example, Barron and Mellow's analysis [2] of May and June 1976 data on a sample of workers unemployed in May revealed that the probability of becoming employed by June was higher for males, those who devoted more time to job search, and those with relatively low reservation wages, and was negatively correlated with receipt of unemployment insurance and length of time unemployed.

While the CPS data on changes in labor force status illustrate the usefulness of longitudinal information, they also illustrate the importance of data accuracy in longitudinal analysis. Woltman and Schreiner [18] have reported evidence that many of the measured gross changes may reflect spurious response changes of persons whose labor force activity has not actually changed. According to monthly average gross flow data for 1977, 48 percent of the CPS unemployed in one month exited from unemployment by the next month. In comparison, when the Census Bureau reinterviewed subsamples of 1977 CPS respondents with regard to the same month, 31 percent of those initially measured as unemployed were measured in the reinterviews as employed or not in the labor force. The high variability in responses for the same period raises the disturbing possibility that many, if not
most, of the measured month-to-month changes in labor force status may be comprised of response changes that would occur even in the absence of any real changes in status. This is not an indictment of the CPS data in particular, but rather a general indication of the sensitivity of flow data to measurement error and of the special importance of data accuracy when addressing the more delicate questions often asked of longitudinal data.

Another example of the use of longitudinal data for change measurement is research on earnings mobility. Cross-section data can reveal what proportion of workers receives low earnings at a point in time, but to measure how many of these low earners stay low earners and how many leave low-earnings status requires longitudinal information. Lillard and Willis' study [9] of PSID data examined the persistence of low-earnings status among white and black men. They defined low earnings in a given year as earnings less than half the median earnings of male workers in the CPS. They concluded that, of the low-earning men in a given year, about 45 percent of the whites and 65 percent of the blacks would still have low earnings the next year. McCall's study [10] of CWHS earnings records obtained roughly similar results. The similarity of the results from both survey and administrative data demonstrates how the validity of one study can be assessed by comparison with another.

A recurring question in analyses of change or persistence in economic status is whether the observed degree of persistence is due to "population heterogeneity" or "state dependence." For example, Plant's study [15] of welfare dependence asked the important policy question of
whether the tendency of welfare families to stay on welfare occurs simply because the same factors that cause them to go on welfare keep them there or whether, in addition, the experience of receiving welfare has some sort of addictive effect that induces continuing welfare dependence. It is usually very difficult to distinguish these two types of processes because their empirical manifestations are so similar. In Plant's study, however, separation of heterogeneity and state dependence was facilitated by the availability of information on both the experimental and the control families in the Seattle-Denver negative income tax experiment. He concluded that the evidence of an addictive state-dependence effect was weak at best. He also discovered that, if he had used arbitrary statistical assumptions commonly employed in analyses of nonexperimental data, he would have been misled into estimating a much larger state-dependence effect. Despite the difficulty of separating heterogeneity and state dependence, researchers have continued to use longitudinal data to address this important issue in such areas as labor force participation decisions and unemployment.3

Analysis of Intertemporal Relationships

Longitudinal data are used not only to measure change in individuals' status over time, but also to relate individuals' experiences or behavior at one time to other experiences or behavior at another time. For example, an individual's early labor market experience might affect his earning capacity in later years, or participation in various government
programs might affect subsequent economic status. Of course, research on such intertemporal relationships requires information on the same individuals at different points in time, i.e., longitudinal data.

One such use of longitudinal data is Ellwood's study [4] of the impact of teenage unemployment on later wages. He analyzed NLS data on young men that finished school between 1965 and 1967 to relate their work experience in their first four years out of school to their wage rates in the immediately following years. He concluded, "Early work experience has a sizable impact on wages. Controlling for individual effects, experience in the second, third, or fourth year out of school tends to be associated with wage increases of between 10 and 20 percent a year."

Another example is Ashenfelter's study [1] of the effect of federal training programs on the later earnings of program enrollees. His sample included 1964 participants in Manpower Development and Training Act programs as well as a comparison group of nonparticipants. He compared the two groups' CWHS earnings records from 1961 to 1969 to estimate the earnings impact of program participation. He concluded that training did increase participants' earnings. He estimated that men's annual earnings were raised, on average, by $150 to $500 in the period immediately following training and by about half as much after five years. For women, the effect appeared to lie between $300 and $600 and did not decline over time.
Analysis Controlling for Unobserved Variables

The third use of longitudinal data is in analysis controlling for unobserved variables. Often in empirical cross-section research, the goal is to estimate the effect of a variable X on a variable Y, holding other variables constant. Frequently, however, some of these other variables either are very difficult to measure or simply happen not to have been collected in the data base. The resulting omission of these unobserved variables from the analysis may bias the estimation of X's effect on Y.

An example is research on the wage effects of union membership. Cross-section studies have compared the wage rates of union members and nonmembers with seemingly similar characteristics and have found that union members generally receive higher wages. Critics of these studies, however, have argued that union members and nonmembers may differ in ways not observable to the researcher. It could be that the union members, even if they had not been in unions, would have earned more than the nonmembers. Although the cross-section studies typically do control for years of schooling, work experience, and other measurable factors, the possibility remains that the estimated union-nonunion wage differential is due to other unobserved factors.

Longitudinal data provide a way of controlling for these unobserved factors. If the effects of unobserved characteristics of the workers stay roughly constant over time, one can estimate union wage effects by examining how the same worker's wages change when he changes union status. If workers typically experience wage gains when they become union members and
wage losses when they become nonmembers, a positive union impact on wages will be estimated. This estimation approach implicitly controls for unobserved fixed effects specific to individual workers by focusing on wage changes of the same workers over time.

Mellow [11] used this type of approach in his study of longitudinally matched CPS data for two samples, one followed over 1974-75 and the other over 1977-78. He found that union membership is associated with about a 7 percent wage premium, smaller than typically found in cross-section studies, but still significantly greater than zero. Mincer [14] conducted a similar study with NLS and PSID data on white males and obtained similar results.

The longitudinal union-nonunion wage studies illustrate some of the pitfalls of longitudinal analysis, as well as its advantages. First, the longitudinal approach may not necessarily eliminate omitted-variables problems. Union joiners or leavers may differ in systematic ways from individuals whose union status does not change. For example, some individuals might become nonmembers because they have been promoted to supervisory positions. For these individuals, union leaving is correlated with wage gains due to a factor other than changed union status. Other individuals may lose union membership because they are laid off from union jobs. These individuals may undergo wage losses due largely to the layoff experience rather than to the change in union status. Recognizing that such factors, if omitted, might bias the estimated wage effect of union membership, Mincer separately analyzed the wage changes of union joiners
and leavers, those who stayed union members, and those who stayed nonmemb-
er members among those that had quit their jobs and those laid off. His results, it turned out, were not dramatically altered. In the PSID data, for instance, relative to job stayers who stayed nonmembers, job quitters who stayed nonmembers experienced an average wage gain of 9 percent, and job quitters who stayed members gained 10.6 percent. In contrast, job quitters who became members gained 17.2 percent, and job quitters who became nonmem-
bers gained only 0.4 percent. These results — showing especially large wage gains for union joiners and especially small gains for union leavers — remain consistent with the finding of a positive union-nonunion wage differential.

A second problem is response error. Even in a cross-section analy-
sis, misclassification of individuals with respect to their union mem-
bership status tends to obscure whatever wage differences actually exist between union and nonunion workers. According to the standard econometric analysis of measurement errors in an independent variable in a regression analysis, the resulting bias in the variable's coefficient is proportional to the ratio of the measurement error variance to the sum of the measure-
ment error variance and the true population variance of the variable. In a longitudinal regression analysis, where change in a variable is the inde-
pendent variable, the bias from response error may be worse for two reasons. First, the measurement error variance may be greater because a response error in either of two periods can cause an erroneous measure of change. Second, the population variance of change in a variable is typi-
cally smaller than the cross-sectional variance in the level of the variable."

In the case of change in union membership status, there is indeed reason to suspect considerable measurement error. Mincer noted that a disturbingly high proportion of those reporting changes in union membership status also reported that they did not change jobs. Suspecting that many of these job stayers had not actually changed union status, he estimated separate wage effects for job stayers and movers. The mover results — such as the ones mentioned above on workers that quit or were laid off — showed more distinct union wage effects than did the stayer results, which probably were biased toward zero by response error. Similarly, Mellow found virtually no union effect among workers that did not change occupation or industry. These results highlight the need to give careful attention to response error when analyzing longitudinal data, especially if the data were obtained in surveys. They also demonstrate the additional care in data collection that may be necessary to obtain answers to the more subtle research questions posed of longitudinal data.
Summary

Recent years have witnessed significant growth in the availability of longitudinal data on labor force experience and behavior. These data—which follow the same individuals over time through surveys, administrative records, or social experiments—have proven extremely valuable for three types of research: measurement and analysis of changes in individuals' status over time (e.g., changes in employment status or income); analysis of intertemporal relationships (e.g., between participation in government training programs and later economic success); and analysis that must control for unobserved variables (e.g., the analysis of union-nonunion wage differences). In some cases, the existence of longitudinal data has opened up avenues of research that simply could not have been pursued otherwise. In other cases, longitudinal data have enabled the examination of previously untestable analytical assumptions and consequently have increased the reliability of research findings.

Despite its great advantages, longitudinal analysis also involves a special problem. Many of the questions addressed with longitudinal data are more subtle than those asked of cross-section data, and their analysis is often sensitive to response error. This sensitivity implies first that researchers should attempt to minimize response error in their choice of data bases. In some cases, for example, data from administrative records may be more accurate than survey data. In addition, longitudinal analysts should examine their data for evidence of response error and explore how response error might affect their results. The sensitivity of longitudinal
analysis to response error also raises the question of whether longitudinal data collection efforts ought to devote more resources to the reduction of such error.

The overwhelming usefulness of longitudinal data for the analysis of many issues has been established by a continuing succession of valuable studies. Because collection of longitudinal data is still a relatively new endeavor, though, several issues associated with their collection need exploration. One important question is how to weigh the sample attrition problems of continuing an old panel survey against the advantages of following the panel over a longer period as well as the large fixed costs of initiating a new survey. Similar questions pertain to the choice between retrospective and repeated interviews. Finally, there exist important and unexplored tradeoffs in allocating survey resources between interviewing more individuals and improving the accuracy of data on those that are interviewed. Some analysis and perhaps even purposive experimentation with alternative approaches to these issues should make the longitudinal data developed in the next decade even more valuable than those of the past decade.
Footnotes


2 Miller [12], Miller and Paley [13], and Fritzker and Sands [16].

3 See, for example, Heckman and Willis [6], Heckman and Borjas [5], and Ellwood [4].

4 This point is developed by Taubman [17].
References


