Imputing income in the Consumer Expenditure Survey

By means of a model-based approach, problems are examined relating to nonresponse in the collection of data on income; new approaches and how they work are set forth

Geoffrey D. Paulin
and
David L. Ferraro

Geoffrey D. Paulin is an economist with the Division of Consumer Expenditure Surveys, Bureau of Labor Statistics. David L. Ferraro is a mathematical statistician with the Demographic Statistical Methods Division, Bureau of the Census.

In the Consumer Expenditure (CE) Survey—the only U.S. government survey that relates expenditures of consumers in the United States to demographic characteristics—income is the characteristic most often used in research and analysis. Indeed, nearly 90 percent of respondents to a recent survey of CE Survey data users reported using the data by income class.[1]

Income data are used in two ways: as a way to classify and as a demographic characteristic. Consumer units [2] are classified by income quintile and level of income before taxes in tables published routinely from the CE Survey. Many users of the data analyze the relationship between income as a demographic characteristic and expenditures. For example, John Sabelhaus used the data to assess the impact of taxing consumption versus income,[3] and William S. Reece estimated the relationship of income to different types of charitable contributions.[4] Other researchers have used income data from the survey to compare expenditure patterns across various groups[5] and to estimate linear Engel curves.[6]

Household surveys, such as the CE Survey, are subject to nonresponse and underreporting. Some respondents provide information on some, but not all, sources of income or on some, but not all, members of the family. The respondent does not always know detailed information on family income, and sometimes the respondent refuses to provide information. In all such cases, the families are classified as nonrespondents, at least with regard to the particular member or source not reported. Underreporters are families who answer the questions, but reduce the real amount. E. Raphael Branch (p. 47, this issue) evaluates the underreporting of income and expenditures.

This article addresses nonresponse to questions about income. Methods of adjusting for nonresponse have ranged from simple account balancing (that is, eliminating families with large gaps between income and expenditures) to selecting *complete income reporters* (described subsequently). Recently, it was decided that a more sophisticated approach was needed to improve income estimates. The Bureau of Labor Statistics and the Census Bureau conducted research jointly to explore how model-based income imputation could be used.[7] The focus in the next few sections is on developing models to estimate income data that are missing due to nonresponse. These models will be evaluated and adapted in the next phase of our research to adjust for underreporting.

The purpose of this article is to describe work that has so far been completed to achieve our goal of imputing missing income using a model-based approach. Experimental data using imputed income values are forthcoming.

Background

Data for the CE Survey have been collected since 1888, when the survey results were used to provide information for tariff negotiations between the United States and European countries. Results are available from the 1901, 1917–19, 1935–36, 1941–42, 1950, 1960–61, and 1972–73 surveys, and from surveys conducted annually beginning in 1980.[8]

From 1901 through the 1960–61 period, the survey was an *annual recall survey*: respondents were asked to recall their incomes and expenditures of the past year. These surveys used balancing criteria to ensure the quality of their data. Families visited had "a disposition to give exact information, and by careful questioning and checking income and expenditures with surplus or deficit, the agent was able to secure a statement that both he and the informant believed to be a very accurate and true presentation of the various expenditures of the family," according to a report to the Commissioner of Labor describing the 1901 survey results.[9]

In the 1935–36 survey, balancing criteria were used to determine whether the family just interviewed would be included in the sample. A completed schedule was rejected if expenditures exceeded income by an unacceptable amount—5 percent for city and village families, and 10 percent for farm families. However, when the time came to plan the 1972–73 survey, questions arose concerning the annual recall approach. Research had shown that better data could be obtained by more frequent interviewing and the use of diaries to account for more frequent purchases.

After much consideration, the quarterly interview and diary components of the CE Survey were implemented in 1972–73. Because data from the interview component were collected quarterly, the balancing criteria were no longer used: they were considered judgmental and too cumbersome for a quarterly survey.[10] Dropping the balancing criteria for a quarterly survey also can be justified by economic theory. Data are collected to reflect quarterly expenditures and annual incomes; because even total expenditures may exhibit seasonality, attempting to reconcile quarterly expenditures with annual incomes may be a problem.[11]

For the first time, two categories of consumers—complete and incomplete reporters—were defined. In general, complete income reporters are families for whom an amount of income from at least one major source is reported for at least one member. All other families were incomplete reporters. (The appendix contains a detailed description of rules defining complete and incomplete reporters.)

Although these two categories formed an important preliminary step in addressing the nonresponse problem, the approach was by no means a complete solution. One reason is that a substantial portion of the sample—about 15 percent—still includes incomplete reporters. A second reason is that even complete reporters include families for whom a full accounting of income is not provided. This may explain why the proportion of complete income reporters has remained between about 84 percent and 90 percent of the population since 1980. (See chart 1.) This has occurred as real expenditures[12] exceeded real income before taxes every year for the lowest two income quintiles and, for some years, for the middle income quintile. (See chart 2.) In 1992, even families in the $20,000–$29,999 income range reported average expenditures—$26,071—that were greater than average income—$24,560.[13]

During the 1980's, several ways to solve these problems were discussed. One idea was to devise new rules for the definition of complete income reporters based on differences between expenditures and income (incorporating the old balancing criteria into the definition of a complete reporter). However, when low-income complete reporters were examined, results indicated that a simple solution was not adequate.[14] Based partly on these results, imputation or other methodologies would be explored to resolve the problem.

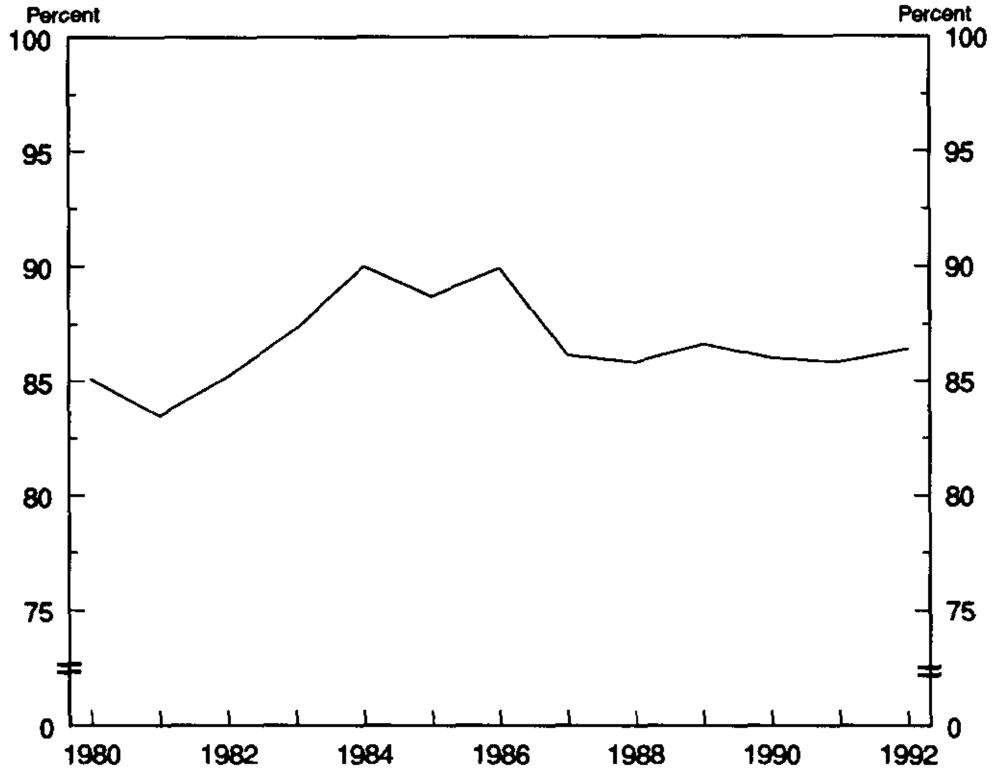
Issues to resolve

Several issues had to be resolved before proceeding with an examination of the various methods. For example, can *hot decking*, a procedure used by some surveys to replace missing values, be used for the CE Survey? If hot decking is not feasible, can *model-based imputation* be used? If so, what problems arise?

Hot decking versus modeling. The Current Population Survey (CPS) uses the hot deck procedure to impute for missing income data collected in its March supplement. The method matches nonrespondents to respondents with similar reported variables, such as age, race, sex, and other demographic characteristics. The respondent's income is then used to replace the nonrespondent's missing value. The number of variables used to match participants varies by survey, but generally, more variables yield better matches than fewer variables.

The problem with using the hot deck method is that the CE Survey sample size is too small. Approximately 5,000 interviews are conducted annually for the survey in which income is collected.[15] This compares with 60,000 families interviewed in the March supplement to the CPS. The smaller the sample size, the more difficult it is to find a suitable donor for a particular nonrespondent. This is reason enough to eliminate

Chart 1. **Complete income reporters as a percentage of the population**



the hot deck procedure as an imputation method for the CE Survey. Other methods need to be explored.

One promising class of methods is *model-based imputation*, described by R. J. A. Little and D. B. Rubin.[16] One member of this class has two major parts. The first requires a statistical model to predict income values. After a suitable model is developed, the second part involves adding an error term to the predicted value to preserve the variance of the distribution.

Model-based methods are useful to examine for several reasons, according to Martin David, the aforementioned Little, Michael E. Samuhel, and Robert K. Triest. These researchers matched CPS data to Internal Revenue Service data to compare methods for imputation. [17] They found that modeling allows "the capacity to include more explanatory variables" than does the hot deck procedure.[18] Although they recommended caution in interpreting their results—limits in matching and comparing variables preclude a definitive conclusion, they wrote[19]—they also found that the model-based alternatives "have slightly lower mean absolute error than the hot deck that is based on the same information."[20] In addition, models "need to be developed to provide realistic competitors to the current hot deck method."[21] But before modeling or analysis of variance can take place, deciding how the data are missing is important.

Response mechanisms. Based on the terminology of Little and Rubin, three types of response mechanism are available: missing completely at random, missing at random, and nonignorable nonresponse.

If the data are missing completely at random, the probability of nonresponse has no relationship with any characteristic of the respondent—every person has an equal probability of refusing. If the data are missing at random, characteristics of the respondent have a relationship with the probability of nonresponse. (For example, older persons may be more likely to respond than younger persons, or homeowners may be more likely to respond than renters.) However, the level of income and probability of nonresponse do not have a relationship once other characteristics are controlled. If the data exhibit nonignorable nonresponse, the level of income is linked with the probability of nonresponse. (For example, persons with high incomes may be less likely to respond than persons with middle incomes.)

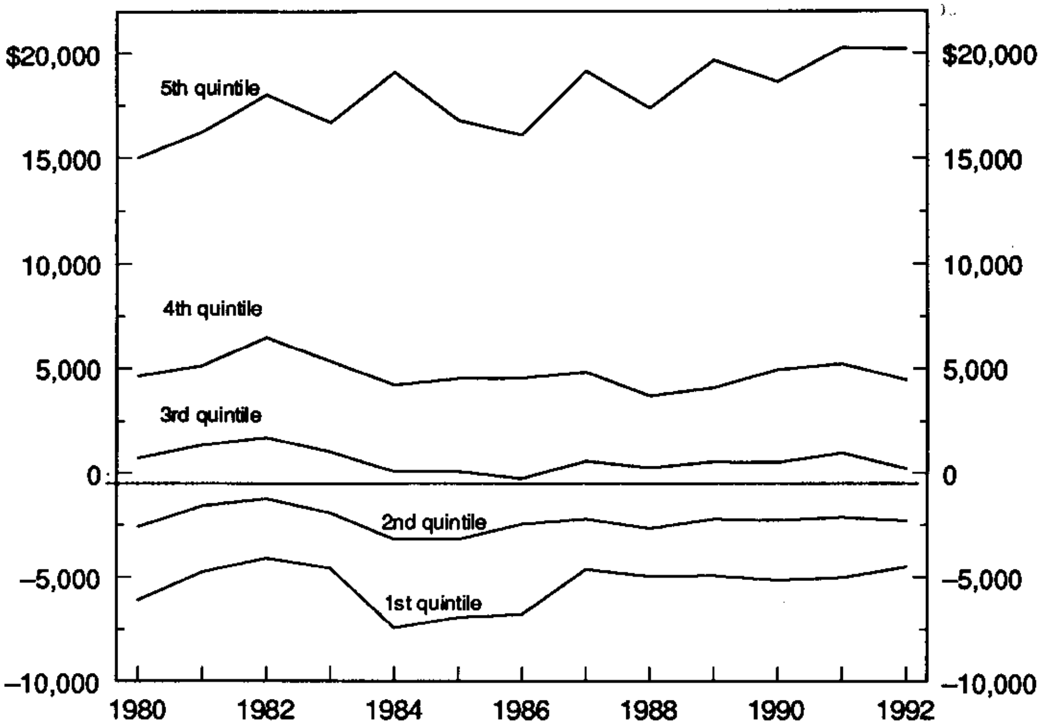
Each of these response mechanisms has implications for what kind of strategy, if any, should be used for imputation. For example, if the data are missing completely at random, there is no meaningful difference between the average income estimated from the sample of those who report income and the true average of the entire population, assuming a large sample size. If the data are missing at random, the average income estimated from the sample will differ from the true average. However, the difference can be explained by differences in the distribution of other characteristics, and imputation procedures are straightforward. If the data exhibit nonignorable nonresponse, the average income estimated from the sample will differ from the true average, and imputation procedures become more complicated. (An example of each case is shown in the appendix.)

After much consideration at BLS and the Bureau of the Census, several reasons were cited for an imputation approach based on missing-at-random assumptions to be explored first. J. S. Greenlees, W. S. Reece, and K. D. Zieschang found evidence in the CPS that income data exhibit nonignorable nonresponse.[22] But later work by Sybil Crawford[23] with the same data found evidence that the missing-at-random assumption was not rejected, partly because the Greenlees, Reece, and Zieschang models were underspecified—that is, they included too few explanatory variables for their model to make accurate predictions.

Additionally, David, Little, Samuhel, and Triest, who also used data from the CPS, found that even though there may be evidence of nonignorable nonresponse in the data, its existence in practice is not qualitatively important.[24] Finally, models based on missing-at-random assumptions provide a baseline of comparison for work based on assumptions of nonignorable nonresponse, which are extremely difficult to model.

Modeling decisions. Two primary issues must be resolved before a model can be developed to predict income. The first is whether the model will be at the member level or family level. The second is whether the model should predict total income or separate income components, such as wage and salary income, self-employment in-

Chart 2. **Difference between incomes before taxes and total expenditures by quintile of income, in 1982–84 dollars**



come, and income from interest and dividends, that are summed to total income.

If the model is at the member level, each member's income is predicted separately, based on individual characteristics such as age and number of hours worked each year. The ease of associating characteristics with earners is the primary advantage of member-level prediction, compared with family-level prediction. For example, at the family level, should the characteristics of the reference person (the first member mentioned when the respondent is asked to "start with the name of the person or one of the persons who owns or rents the home") be used, or those of someone else?

Economic theory suggests that interactions take place between family members when work decisions are made. For example, how much member A works may depend on how much member B earns. If the model is at the family level, member-level interactions are no longer of concern, because they already have occurred and the outcome is observed. For these reasons, we have explored family-level incomes first.

With regard to the second issue, the primary advantage of predicting total family income is simplicity: only one model is used, and a summation of components is not necessary. However, information on the composition of income and level of components is lost to the researcher if only total income is imputed. Furthermore, some researchers may argue that individual sources of income should be modeled separately because labor income—wages and salaries—and nonlabor income—interest and dividends—are predicted by models that require different variables. Even in cases where some variables are expected to be statistically significant predictors of income in both models (for example, age), the parameter estimates are almost certainly different across models. For these reasons, we have examined components of income first.

Research

Modeling income. Crawford [25] launched preliminary investigations into modeling wage and salary incomes for single persons by means of ordinary least squares regression[26] with a small number of independent variables. However, Geoffrey D. Paulin and Elizabeth M. Sweet conducted more extensive work with model-building in which they attempted to find predictive independent variables that Crawford and others had not considered.[27] The work by Paulin and Sweet also is more complicated because it focuses on larger (two-member) families and calls for additional statistical techniques. For example, the authors use two approaches to model wage and salary incomes and merge the results.

The first approach uses ordinary least squares and stepwise regression.[28] The final model is run with weighted least squares[29] to correct for heteroskedasticity (a condition in which the data do not vary uniformly around the regression line). The dependent variable (wage and salary income) is not transformed throughout. The second method derives complex interaction terms for use as independent variables. In this stage, the authors use the natural logarithm of the dependent variable to correct for heteroskedasticity. The resulting models from the two methods are then merged, and final variables are selected, using the combination of ordinary least squares and stepwise regressions, as in the first method.

The procedures are run using both weighted least squares and the natural logarithm as a dependent variable. In each case, some variables selected from the first method and some selected from the second remain in the final, merged, model. The procedures (weighted least squares and the natural logarithm as dependent variable) are then tested to determine what specification the final model should have. Although the model using the natural logarithm has a lower mean square error when predicted versus actual wage and salary incomes are compared,[30] Paulin and Sweet find the smallest mean square error when they subject wage and salary data to a Box-Cox transformation, the formula for which is

$$Y^* = (Y^\lambda - 1)/\lambda,$$

where Y is the variable to be transformed (wage and salary income in this case), λ is a parameter whose value is found by experimentation, and Y^* is the resulting variable, which is then used in regressions in place of Y.

Expenditures and income. In the next step of research, Paulin and David L. Ferraro [31] explored whether expenditure data are useful in predicting income. The authors examined wage and salary and self-employment incomes separately for single persons. They compared models with no expenditure data, total expenditures, and selected expenditures, such as for food at home, shelter and utilities, and telephone services. Box-Cox transformations were calculated for all income and expenditure data.

Paulin and Ferraro found that each of the expenditures examined has some predictive power compared with no expenditures, but that total expenditures produce the model with the largest R^2 value.[32] Among the less aggregated expenditures, basic goods and services (food at home, shelter and utilities, and apparel and services) add the most to the model's predictive power, as measured by R^2, while food at home adds the least. Paulin and Ferraro also showed that income shares—the level of the specific expenditure divided by income—are not very useful in predicting income.[33] They plan to continue the research by examining families to decide whether expenditures should be used to predict incomes.[34]

Imputation using the models. Nanak Chand and Charles H. Alexander [35] have described how the final income model can be used to yield imputed values. Their method imputes income for multiple-member families with multiple income sources. At this point, they have developed a stochastic method that imputes income for each member and each source of income separately. Employing a stochastic method is essential if valid inferences are to be drawn from the imputed data.[36]

In Chand and Alexander's procedure, random variables are generated with replacements that are used as the imputed values. The imputation accounts for the variability of the observed values of the variables and preserves the observed relationships between family members and sources of income. In the original approach, based on Little and Rubin, a vector of imputed values is drawn from the multivariate normal distribution in which parameters are estimated from the data on income reporters. More recent work has focused on imputing the variables sequentially, at each step using previously imputed values as independent variables for the next step. The sequential approach is more readily applied to various household sizes and patterns of income sources and easily extends to nonnormal distributions.

Related and future work

Other researchers are exploring issues related to nonresponse to questions about income. For example, J. L. Eltinge and I. S. Yansaneh[37] have studied weighting adjustment as a correction for such nonresponse. Under the missing-at-random assumption, weighting can eliminate the bias due to nonresponse in estimates of mean income. The method works by accounting for a percentage of persons in a particular group who do not respond to questions about income and who will receive a value of zero for their incomes. Using the inverse of the probability of responding to develop a new weight for adjusting the mean should provide an estimate of the "true" mean for the group. In its simplest form, an example of an adjustment would be to multiply the mean of the sample (respondents and nonrespondents, where nonrespondents are given a value of zero) by $1/X$, where X is the proportion of people who provide an answer. Eltinge and Yansaneh discuss in detail finding the proper value for X and how to use X properly in the weighting adjustment.

Other work in progress explores additional methods of testing the missing-at-random and nonignorable nonresponse assumptions, methods for modeling income if nonignorable nonresponse is found, and methods for final implementation if it is decided to impute income under assumptions of nonignorable nonresponse.

Finding an appropriate method of adjusting for nonresponse is an arduous and complicated process. However, much progress has been made in the last few years. Through continuing research, CE income data are better understood and are therefore better able to be adjusted. Plans are in the works to make available, through various media, the experimental imputed income values to allow researchers to use these data and explore the effects of imputation on their results.

Footnotes

[1] *Consumer Expenditure Data User Survey Results*, Report 832 (Bureau of Labor Statistics), March 1993, p. 2.

[2] A consumer unit is the basic unit of analysis in the CE Survey. It is defined as a single person living alone or sharing a household with others, but who is financially independent; members of a household related by blood, marriage, adoption, or some other legal arrangement; or two or more persons living together who share responsibility for at least two of three major types of expenses—food, housing, and other expenses. The terms *family* and *families* are substituted throughout for convenience.

[3] John Sabelhaus, "What Is the Distributional Burden of Taxing Consumption?" *National Tax Journal*, September 1993, pp. 331–44.

[4] William S. Reece, "Charitable Contributions: New Evidence on Household Behavior," *American Economic Review*, March 1979, pp. 142–51.

[5] Older workers and nonworkers (Thomas Moehrle, "Expenditure Patterns of the Elderly: Workers and Nonworkers," *Monthly Labor Review*, May 1990, pp. 34–41); single- and dual-earner families (Rose M. Rubin, Bobye J. Riney, and David J. Molina, "Expenditure Pattern Differentials between One-Earner and Dual-Earner Households: 1972–73 and 1984," *Journal of Consumer Research*, June 1990, pp. 43–52); and ethnic groups (F. N. Schwenk, "Income and Consumer Expenditures of Households Headed by Hispanic and Elderly Women," *Family Economics Review*, 1994, pp. 2–8).

[6] Barbara A. Sawtelle, "Income Elasticities of Household Expenditures: A U.S. Cross-Section Perspective," *Applied Economics*, May 1993, pp. 635–44.

[7] The Census Bureau collects the data under contract with the Bureau of Labor Statistics, which conducts the survey.

[8] For an analysis of historical spending patterns from 1901 through 1986–87, see Eva Jacobs and Stephanie Shipp, "How family spending has changed in the U.S.," *Monthly Labor Review*, March 1990, pp. 20–27.

[9] *Eighteenth Annual Report of the Commissioner of Labor, 1903: Cost of Living and Retail Prices of Food*, Washington, DC, pp. 16–17.

[10] For more on the methodology and results of earlier surveys, see Eva Jacobs and Stephanie Shipp, "A History of the U.S. Consumer Expenditure Survey: 1935–36 to 1988–

89," *Journal of Economic and Social Measurement*, 1993, pp. 59–96.

[11] For a description of seasonality with regard to the data in the CE Survey, see the article by Moehrle, pp. 38–46, this issue.

[12] For example, total expenditures reported in the interview component of the CE Survey divided by the Consumer Price Index, to control for price changes.

[13] These numbers are taken from the integrated survey results, which combine results from the diary and interview components of the survey. (See appendix.) Although the interview component is the focus of this article, integrated results are cited because annual data from the CE Survey are published in the integrated format and the income data used for each group are derived from the interview component in any event. (Integrated data are not used in chart 2 because they are not available before 1984.) Although 1992 total expenditures ($24,329) for the $20,000–$29,999 income group, as reported in the interview component, are less than average income before taxes, the difference is only $231, indicating that the missing data are still important.

[14] See Geoffrey Paulin, "Income Project Results," internal memorandum, Bureau of Labor Statistics, Division of Consumer Expenditure Surveys, July 24, 1989. Paulin explored the possibility of redefining complete reporters with income of less than $5,000 as incomplete reporters if their total expenditures exceeded their income by at least 20 percent. However, even when families who might legitimately have such a deficit (for example, if the reference person was retired, a student, or self-employed) were retained in the complete-reporting group, 38 percent of the sample would have been redefined as incomplete reporters under the proposal.

[15] That is, there are about 5,000 families who have a second interview over the course of a year. Although income data also are collected in the fifth interview, data collected during the second interview have been the primary focus. Fewer problems arise with attrition from the second interview, and fewer problems should occur with sample selection bias from the second than from the fifth interview. Sample selection bias is a problem for determining response mechanisms, as described in the text, if the probability of dropping from the sample is related to the same characteristics as nonresponse.

[16] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data* (New York, John Wiley and Sons, 1987).

[17] Martin David, Roderick J. A. Little, Michael E. Samuhel, and Robert K. Triest, "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association*, Applications, March 1986, pp. 29–41, especially pp. 39–40.

[18] *Ibid.*, p. 40.

[19] *Ibid.*, p. 29.

[20] *Ibid.*, p. 40.

[21] *Ibid.*, p. 29.

[22] J. S. Greenlees, W. S. Reece, and K. D. Zieschang, "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association*, June 1982, pp. 251–61.

[23] Internal memoranda, Bureau of Labor Statistics, 1989 through 1990.

[24] David, Little, Samuhel, and Triest, "Alternative Methods," p. 40.

[25] Internal memoranda, Bureau of Labor Statistics, 1989 through 1990.

[26] Ordinary least squares is a technique in which the sum of the squared deviations of data points from a regression line is minimized.

[27] Geoffrey D. Paulin and Elizabeth M. Sweet, "Modeling Income in the U.S. Consumer Expenditure Survey," 1993 *Proceedings of the Section on Research Survey Methods*, Vol. 1, American Statistical Association, pp. 98–106.

[28] In stepwise regression, the F-statistic for each variable, if it is included in the model, is calculated, and then the variable with the highest F-statistic is added first, followed by the second, and so forth, assuming that the F-statistic meets the set value for inclusion in the model. At each step, the F-statistics for all variables in the model are recalculated. If any fall below the required value for remaining in the model, they are removed from the model.

[29] Weights are calculated from regressions of predicted wage and salary incomes on residuals.

[30] Predicted values of the natural logarithm are exponentiated and compared with observed wage and salary income before the mean square error is calculated.

[31] Geoffrey D. Paulin and David L. Ferraro, "Do Expenditures Explain Income?" paper presented at the Joint Statistical Meetings of the American Statistical Association, August 17, 1994, Toronto, Canada.

[32] R^2 is defined as the ratio of the variance explained by the model to the total (that is, explained plus unexplained) variance. As R^2 approaches 1, the percent of the total variance explained in the dependent variable (wage and salary income in this case) approaches 100.

[33] The models that use income shares as a dependent variable have low R^2 values and predict a high percentage of negative shares.

[34] Paulin and Ferraro state some concerns about using expenditure data, in addition to the usual concerns with statistical significance. The primary issue is whether endogeneity is a problem for researchers using the data. For example, if food at home is used to predict income, then users who are trying to estimate the income elasticity of food at home may have problems with endogeneity. However, if the prediction of income is strong enough, these concerns may be less important. Therefore, further exploration is necessary.

[35] Nanak Chand and Charles H. Alexander, "Imputing Income for an N-Person Consumer Unit," paper presented at the Joint Statistical Meetings of the American Statistical Association, August 15, 1994, Toronto, Canada.

[36] See Little and Rubin, *Statistical Analysis*.

[37] J.L. Eltinge and I.S. Yansaneh, "Weighting Adjustments for Income Non-Response in the U.S. Consumer Expenditure Survey," Technical Report No. 202 (Bureau of Labor Statistics, 1993).

APPENDIX: About the survey

The Consumer Expenditure (CE) Survey is the most detailed source of data on family expenditures collected by the Federal Government. The survey collects information on demographics and income for members of the family, providing an important resource for researchers in economics, marketing, policy analysis, and related fields.

The quarterly interview and the diary are the two components of the CE Survey. The quarterly interview is a panel survey to collect information on expenditures from families over five consecutive quarters. During each interview, the respondent is asked to recall expenditures for the last 3 months for most items in the survey. The first interview is to ensure that expenditures reported took place in the time frame in question.

The interview component is designed primarily to collect recurring expenditures, such as rent, and "big ticket" expenditures, such as automobiles or major appliances, because outlays for such items tend to be remembered for long periods.

Families participating in the diary component receive a diary for 2 consecutive weeks to record their expenditures during the survey period. The diary component is designed to collect expenditures for frequently purchased items, such as groceries, and small-cost items, such as laundry detergent.

Although each component collects information on income, the interview component is examined in this article for several reasons. First, income data reported in the interview component are generally considered to be more reliable than income data reported in the diary component. This is because, although the questions in both components are similar and are each asked by an interviewer, the questions in the interview component are asked at the end of the interview, after the demographic questions and the questions on expenditures have been asked. Thus, respondents may be more willing to answer questions on income after answering other questions and may be less hesitant to answer questions to which they might otherwise be sensitive.

Second, the percentage of incomplete reporters of income is generally larger in the diary component than the interview component, probably for the reasons just described. (Incomplete reporters usually comprise about 15 percent of the latter and about 20 percent of the former.)

Third, the interview component collects information on about 95 percent of all expenditures, making it more useful for general analyses, although the diary component is a better source for certain expenditures, such as detailed food items.[1] Most studies, including those cited at the beginning of this article, use the interview component, the most common exception being detailed food analyses.

Income data are collected in the interview component during the second and fifth interviews. Information on some sources is collected for each member of the family who is at least 14 years old, and information on other sources is collected for the entire family. (See below for a complete listing of sources of income and the level of detail at which information is collected.) In each case, the respondent is asked to recall the information for the year before the interview. Data on total family income (the sum of income from all sources for each member and for the entire family) are published in several forms. Member-level income data are made available to researchers.

Related questions are asked, such as questions about the number of hours per week and weeks per year worked, the occupation of the jobholder, and the type of work performed (wage and salary work or self-employment). Detailed demographic information on each member of the family (age, race, sex, and educational attainment) also is collected. Where appropriate, characteristics such as age or education are updated during each interview.

Sources of income. The following list describes what data are collected at what level of detail:

Collected at the member level
- wage and salary;
- self-employment;
- Social Security and Railroad Retirement benefits;
- supplemental security income.

Collected at the family level
- unemployment compensation;
- workers' compensation and veterans' benefits;
- public assistance;
- interest on savings accounts and bonds;
- regular income from dividends, royalties, estates, or trusts;
- pensions or annuities from private, military, or other government sources;
- net income or loss from renters or other payments received;
- regular contributions for support, such as alimony and child support;
- money income from care for foster children, cash scholarships, and fellowships or stipends not based on working;
- food stamps.

Complete and incomplete income reporters. Families that fit at least one of the following criteria are classified as complete reporters:

1. All major sources of income for each member are reported as zero or a valid blank, and at least one member reported a valid, nonzero value for another source of income.
2. The reference person reported zero or valid blanks for all major sources of income, and at least one other member reported a valid, nonzero amount for at least one major source of income.
3. The reference person reported a valid, nonzero amount for at least one major source of income.

Valid blanks result when there is a good reason to leave a question unanswered. For example, if a member of the family did not work at all during the past year, then a valid blank appears for that member's salary earnings. For some sources (for example, self-employment income), negative amounts can be valid responses.

A family whose reference person reports a major source of income is classified as a complete income reporter, even if other members do not have valid responses. But if there are no valid reports for major sources for the reference person, the family is classified as an incomplete reporter of income, even when all other members have valid responses.

Response mechanisms. To illustrate the implications for average income of the three assumptions discussed earlier, the following equations are given:

M_i = income for the *i*th male.
F_i = income for the *i*th female.
$M = \Sigma M_i; F = \Sigma F_i; M \neq F.$

a, b $(a \neq b)$ are parameters describing the percent of respondents reporting income.

m = total number of males in the sample;
f = total number of females in the sample.

If there are no missing values, then the true average income can be computed directly to be

$$(M + F)/(m + f).$$

Under the assumption that the data are missing completely at random, males and females have the same probability of reporting. Therefore, the mean of the respondents is

$$\begin{aligned}&(aM + aF)/(am + af)\\&= a(M + F)/a(m + f)\\&= (M + F)/(m + f).\end{aligned}$$

Hence, the mean of the respondents is identical to the true mean. However, if the data are missing at random, the sample mean is computed as

$$(aM + bF)/(am + bf).$$

Because the mean of the respondents is not the same as the true mean, some correction is required. Imputation is straightforward because a method can be devised in which incomes are modeled for men and women, taking into account the difference in average incomes between the two. However, if the data exhibit nonignorable nonresponse, the situation is much more complex. The mean of the respondents then becomes

$$[g(\dot{a}(M_i),M_i) + h(b(F_i),F_i)]/[m(a(M_i)) + f(b(F_i))].$$

In this case, the observed incomes are a complicated combination of the probability of response and level of income. Disentangling these effects is extremely challenging.

Footnote

[1] Although respondents are asked to report all expenditures incurred in the 2-week period in which they participate, few "big ticket" items are collected in the diary component because of the short reporting period and infrequency of purchases of such items.