Journal of Official Statistics, Vol. 12, No. 4, 1996, pp. 403–419

Modeling Income in the U.S. Consumer Expenditure Survey

Geoffrey D. Paulin[1] *and Elizabeth M. Sweet*[2]

Nonresponse to income questions is common in household surveys. Using data from the U.S. Consumer Expenditure Survey, the authors explore different procedures designed to yield a model-based imputation strategy for wage and salary income of two-person consumer units. Selected variables from each are synthesized into a final model that is tested for proper specification. Results of the final model indicate that imputation will increase the means of published Consumer Expenditure Survey income data.

Key words: Missing data; nonresponse; imputation; missing at random.

1. Introduction

Income is one of the most important variables in any study of consumer issues. It can be used to group consumers by purchasing power, or to predict the level of expenditures for a given item. Income elasticities measure the responsiveness of purchases of goods and services to changes in income. Income is also important in determining the probability of purchasing some goods and services. For example, high income families are more likely to hire domestic help or to go out to eat than are low income families. Unfortunately, perhaps because of its importance, income is also one of the most sensitive demographic characteristics collected in many surveys. Often consumers refuse to report any income at all, and many who report some sources of income do not report other sources.

The 1988-90 U.S. Consumer Expenditure Survey, sponsored by the U.S. Bureau of Labor Statistics and collected under contract by the U.S. Bureau of the Census, contains detailed information on family level spending and demographic characteristics. These data are collected during the second through fifth interview in a series of five quarterly interviews, each consisting of about 5,000 consumer units. (See Appendix A for definition.) Income data are collected during the second and fifth interviews only. Data from the first interview are used strictly for bounding purposes, and are not published.

[1] Economist, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE #3985, Washington, D.C. 20212, U.S.A.
[2] Mathematical Statistician, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.
Acknowledgments: The authors wish to thank Roderick J.A. Little, University of Michigan and Brent R. Moulton, U.S. Bureau of Labor Statistics for their comments as discussants of an earlier version of this work, which was presented at the August 1993 Joint Statistical Meetings of the American Statistical Association, San Francisco, California. The authors also wish to acknowledge Charles H. Alexander, Jr., U.S. Bureau of the Census for his guidance, especially in the sections describing Bureau of the Census procedures. This article describes the results of research undertaken by the staffs of the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census. All material contained herein is solely the responsibility of the authors, and does not necessarily reflect the views of either agency.

Currently, consumer units are divided into two groups for publication purposes: "complete" and "incomplete" income reporters, depending on the respondent's answers to income questions. Although 85% of consumer units are classified as complete income reporters, even these families do not always provide a complete accounting of all types of income. (See Paulin and Ferraro (1994) for a detailed description of complete and incomplete reporting definitions, sources of income collected in the Consumer Expenditure Survey, and other background information.) It is hoped that imputing data to replace missing income values will allow more complete usage of the data for both research and publication of Consumer Expenditure Survey data.

This article describes modeling techniques currently under joint investigation by U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census as part of a strategy described by Little and Rubin (1987). Using different techniques, each group works on separate models that are merged at the conclusion of the project. The product is a final model for imputation that includes the best results from both strands of research.

Presumably, missing income from all sources can be imputed with varying success. Wage and salary income is modeled in this article because it is the most frequently reported type of income; about two-thirds of consumer units that are complete income reporters report wage and salary earnings. It is also assumed to be the most accurately reported type of income, because people generally have a good idea of their own (and other members') wage or salary levels.

2. Preliminary Issues

Before deciding on an imputation strategy, several important issues are decided: First, are the income data missing randomly, or is there a pattern to nonresponse? Second, which consumer units should be modeled? Third, should income be modeled at the member level (and then aggregated), or for the consumer unit as a whole?

2.1. Definitions of missingness

According to Little and Rubin (1987), when the response variable (income in this case) is missing, the problem can be classified as missing completely at random, missing at random, or nonignorable nonresponse. If the data are missing completely at random, the probability of nonresponse is constant, and therefore independent of all demographic characteristics of the respondent. Under the missing at random assumption the probability of nonresponse may differ with respect to demographic characteristics, but not with respect to the response variable. If the probability of missingness is related to the level of the response variable, then nonignorable nonresponse conditions hold.

When choosing an assumption about the data, missing completely at random is eliminated immediately. It is a rather strong assumption, at least when applied to income data. It also implies that nonresponse has no effect on mean income – that is, the mean of a large sample will not differ from the population mean due to nonresponse bias (Paulin and Ferraro 1994, p. 31), thus negating an important reason for imputation. For these reasons, the missing completely at random assumption is not considered further for modeling income.

However, the Consumer Expenditure Survey wage and salary income data are assumed to be missing at random. Several factors are important in this decision: First, missing at random assumptions provide a baseline on which future work can be built; it is easier to proceed under missing at random assumptions to identify potential problems that may arise under the more complex assumptions of nonignorable nonresponse. Second, Crawford (1989-90) finds evidence that missing at random is a plausible assumption. Crawford analyzes wage and salary income data from the Current Population Survey that has been previously matched to data from the Internal Revenue Service. Although there are no matched income data available for the Consumer Expenditure Survey, the wage and salary income questions in the Consumer Expenditure Survey are quite similar to those in the Current Population Survey; therefore, Crawford assumes that if the response mechanism is missing at random in the Current Population Survey, the same mechanism operates in the Consumer Expenditure Survey. Although earlier work by Greenlees, Reece, and Zieschang (1982) finds that income data in the Current Population Survey are not missing at random, Crawford (1989) implies that their model may have an insufficient number of background variables.

2.2. *Consumer unit size*

In order to work with the least complex data first, both the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census agreed to start with single-member consumer units and, based on the results, build separate models for two-member and eventually multiple-member consumer units. Because single-member consumer units have few complications, and multiple-member consumer units are the subject of future work, in this article the primary focus is on two-member consumer units. Single-member consumer units are discussed as necessary for illustration.

2.3. *Consumer unit versus member level income*

Consumer unit level income is examined instead of member level income for several reasons. First, the goal is to impute consumer unit income, because expenditures are obtained and published at the consumer unit level. Second, the error in imputing consumer unit income directly is probably less than from summing across imputed member incomes, particularly if incomes are imputed for multiple members. The joint probability distribution between the variables is difficult to preserve in this case. Third, members of the consumer unit are assumed to decide how much to work based on how much nonlabor income (interest, pensions, Social Security, etc.) is available to them individually, or to the consumer unit as a whole. Members also may view each other's incomes (whether from salary or not) as nonlabor income. Trying to capture these interactions at the member level can be difficult even in theory; from a practical standpoint, they are often impossible to capture, because many income sources are collected for the consumer unit as a whole. But at the family level the outcome of such interactions is observed. For these reasons addressing consumer unit level wage and salary incomes provides an important first step in the modeling procedure; member level incomes will be explored in future work.

3. The Models

3.1. Data

The data are from second interviews occurring between the first quarter of 1988 and the fourth quarter of 1990 for two-member consumer units (husband and wife only, a single parent with one child, and other two-member consumer units) in which at least one person reported wage or salary income. Although it is first hypothesized that married-couple consumer units are different from single parent and other families, the Chow test (Kennedy 1992; Maddala 1988, pp. 130–137) does not confirm a statistically significant difference between these two groups, at least not at the 5% significance level. However, the test does react at the 10% significance level. Therefore, the samples are pooled, but separate dummy variables for single parents and for other families are kept in the model. (The Chow test is described briefly in Appendix B.)

Of the 2,793 families initially selected, 50 have missing values for at least one independent variable and are dropped in the regression stage. There are 2,743 families included in the regression results. Sample weights are used in the regressions to reflect the population and to account for sample design effect.

3.2. Variable selection

Because the goal of imputation is to predict income as accurately as possible, the proposed model in theory contains as many independent variables as may be plausibly related to income. Even though increasing the number of variables means increasing the probability of severe multicollinearity, the problem at hand is to predict income values, rather than to find precise relationships between income and independent variables. Therefore, multicollinearity is not so serious an issue in imputation. Little (1993) supports this position. However, to minimize processing costs when the imputation is implemented, both the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census attempt to find models with maximum predictive power and a minimum of variables.

3.3. U.S. Bureau of Labor Statistics procedures

The single-member consumer unit offers the fewest problems in initial variable selection. But for two-person consumer units, some variables are not so obviously chosen. For example, whose age should be chosen – that of the oldest person, the male (if there is one), or someone else? After some consideration and testing, both the age of the principal earner (i.e., the person whose contribution to consumer unit income is the highest) and that of the other person are included.

The full model includes numerous independent variables that are tested for inclusion in a reduced model in the following way: First, the family wage and salary income (that is, the level of wage and salary income reported for the consumer unit as a whole) is regressed on the full set of independent variables using ordinary least squares. At the same time, a stepwise regression is performed on the same set of variables. The results of each are compared. If either procedure finds that an independent variable is statistically significant, that variable is retained for further testing; otherwise, it is removed, unless there is a good

reason from economic theory or applied statistics to keep it. For example, if age is not statistically significant, but the interaction between age and education is, then age is retained. Or if most, but not all, of a group of related dummy variables are found to be statistically significant, all are retained. (For a complete list of all variables considered, see Paulin and Sweet 1995, Tables 1 and 3.)

The steps described above are repeated until a final reduced model emerges. Then the residuals are examined to ascertain whether they are random, or whether they are related to one or more of the independent variables in the model. To test the hypothesis that the residuals are unrelated to other independent variables, the residuals are regressed on the other independent variables in the model. If any coefficient is statistically significant, the hypothesis of random residuals is rejected.

Because the residuals invariably are related to more than one independent variable, the absolute value of the residuals (and squared residuals) are regressed on functions of their associated predicted values, i.e., the level of wage and salary income the model predicts for the consumer unit. Results from the residual regressions are then used to weight the final model. This two-step weighted least squares procedure helps to correct for heteroscedasticity (Maddala 1988, pp. 162, 170–172). When more than one weight looks plausible, a series of tests including the Breusch-Pagan, Goldfeld-Quandt (Maddala 1988, pp. 164–167), and Park-Glejser (Pindyck and Rubinfeld 1981) are used to see which weight appears to reduce the problem the most. (Descriptions of these tests are provided in Appendix B.)

3.4. U.S. Bureau of the Census procedures

The U.S. Bureau of the Census' model uses a semilog specification (i.e., the natural logarithm of a dependent variable is regressed on untransformed independent variables) instead of weighted least squares to reduce heteroscedasticity. (The advantages and disadvantages of each specification are described in Section 4.)

For single-member consumer units there are only a few instances of strong interactions between variables or of variables that need collapsing. But for two-member consumer units, it is not clear how member-level variables should be used or transformed into consumer unit-level variables. Model selection and variable creation occur simultaneously because the initial variables are selected arbitrarily.

Initially, the member-level variables are combined and transformed into a set of consumer unit-level variables that are hypothesized to be related linearly to income. In the process high collinearity, which sometimes results from including each member's characteristics in the model, is avoided.

The ultimate goal is to find a consistent model with as high an R^2 value and as few degrees of freedom as possible. The variables should be approximately orthogonal to each other with respect to income, and make intuitive sense. Some modification to the usual forward selection process is needed because of the lack of a well-defined set of variables relevant to the problem. To achieve this, the "Transformed Main Effects" method is developed.

The Transformed Main Effects method is an offshoot of the forward selection process. The first step of the Transformed Main Effects method is to select the variable that

produces the highest R^2 value among all the categorical variables, an example of which is "occupational class" in Table 1. Next, the LSMEANS (SAS procedure that adjusts means for unbalanced design) are examined. In situations where the categorical variable is not binary, categories are collapsed based on t-tests and plausible, intuitive interpretations to create a final variable, X_1, that is either binary or that has a small number of meaningful categories. With X_1 in the model, the next strongest variable (X_2) is chosen for entry. This is the variable that produces the highest R^2 in a model that includes an intercept, X_1, X_2, the interaction between X_1 and X_2, and an error term (model A). If the interaction term is significant, then LSMEANS are examined for model A minus X_1 and X_2 (model B). If the predictive power of models A and B are identical, the interaction term is treated as a categorical main effect variable, and statistically insignificant categories within this main effect are collapsed again based on t-tests and intuitive interpretation. If the interaction is not significant, model C (i.e., model A minus the interaction term) is examined. The categories in variable X_2 are collapsed based on their LSMEANS and intuitive interpretation, and the categories in variable X_1 are re-examined for changes. This process continues until all variables are tested.

In theory, each step could add to the *model* one degrees of freedom; often, though, this does not happen. For example, one strong interaction term may substitute for the addition of several categorical variables. The result is a model with fewer model degrees of freedom than the initial model, but with a similar R^2. Simultaneously, the effects of multicollinearity are reduced using the Transformed Main Effects method, because the newly created variables are often by definition orthogonal to related categorical variables.

While the Transformed Main Effects method uses a forward selection method in selecting independent variables (as opposed to the stepwise procedure used by the U.S. Bureau of Labor Statistics), both methods select many of the same variables.

4. Merging Models

Once final results from each method are obtained, the next step is to merge them into one model. One important question is whether to use the weighted least squares or semilog specification. The main advantage of weighted least squares is that the parameter estimates can be directly interpreted. For example, if the equation turns out to be $Y_w = i_w + 5A_w$, where Y_w is weighted income, i_w is the weighted intercept, and A_w is weighted age, one can say that income increases 5 USD for every year age increases. In contrast, a similar specification under a semilog model, $\ln Y = i + 5A$, means that the *logarithm* of income increases by 5 for every year age increases, which is a much less intuitive statement, even though the semilog specification has some well-known intuitive properties. For example, if the parameter estimate on age is small in the semilog case, it can be interpreted as the percentage change in income given a unit increase in age. However, of interest here is the change in the *actual value* of the dependent variable Y, not the *percentage change* in Y or even the *actual* change in $\ln Y$; also of great interest is how the *actual* value of Y differs from the *predicted* value of Y. From an intuitive standpoint, it is easier to envision how well a model fits a data point when the predicted Y is 10,000 USD and actual Y is 12,000 USD than it is to envision the fit of a model where predicted $\ln Y$ equals 9.21 and actual $\ln Y$ equals 9.39, because most people do not think in logarithmic terms. Nevertheless, the main advantage of the semilog

model is that it is much easier to use than the weighted least squares method; that is, calculations and tests of various weighting schemes are not necessary with semilog models.

Three experiments are carried out to determine which specification (weighted least squares or semilog) to use. The first step in each case is to take all independent variables from the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census final models and put them into the right-hand side of a regression equation whose dependent variable changes with each experiment about to be described. The second step is to carry out a series of tests to ascertain which model performs best.

4.1. Experiment 1

Two models are run. The first uses family wage and salary income as its dependent variable, as the U.S. Bureau of Labor Statistics does. The second uses the natural logarithm of family wage and salary income as the U.S. Bureau of the Census does. The U.S. Bureau of Labor Statistics procedure is followed until a reduced model emerges in each case. Residuals from the salary model are then tested and an appropriate weight is found so that weighted least squares can be run on this reduced model to get final results.

The next step is to decide which reduced model – weighted least squares or semilog – produces the best results. To do this, the predicted values from the weighted least squares model are calculated. The predicted values from the semilog model are exponentiated to convert them to salary estimates. These values are used to calculate error terms for each observation. Then the squared error terms are summed. This grand total is divided by the number (n) of consumer units in each regression model, yielding a mean squared error.

Because the mean squared error from the weighted least squares model (487,994,732) is larger than the mean squared error from the semilog model (479,931,419), the semilog model yields better results. This may be because the semilog model never allows a negative prediction for total wage and salary income. The weighted least squares technique, however, sets no lower bound on predicted income, and indeed some negative wage and salary incomes are predicted. Because these would be set to zero anyway if this method were used for imputation, all negative predicted values are converted to zero and the same procedure is followed. Although the weighted least squares numbers improve (i.e., the mean squared error drops to 486,275,951), the semilog model still appears to be the better approach. Another possible explanation for the superiority of the semilog in this case arises from a subtlety implied by the semilog specification. If the true relationship between income and characteristics is semilog, then

$$E(Y|X) = E[\exp(X\beta + \varepsilon)|X] = \exp(X\beta)E[\exp(\varepsilon)]$$

However, the equation is true only when $E[\exp(\varepsilon)]$ equals one. If the true relationship between income and characteristics is not semilog, then taking the antilog in this way purges some of the error, thus making the semilog model perform better in this experiment because the semilog model yields biased error terms.

4.2. Experiment 2

The independent variables from the reduced weighted least squares and semilog models calculated in Experiment 1 are merged in the same way as the final U.S. Bureau of Labor

Statistics and U.S. Bureau of the Census models are merged. But now Bera-McAleer and PE tests (Maddala 1988, pp.179–180) are used to indicate which approach might be preferred.

At first both tests are adapted for weighted least squares by using the weighted least squares predicted values instead of unweighted predicted values. Unfortunately, the results of both tests are ambiguous. To make sure that the adapting of the tests for weighted least squares is not the problem, the same tests are run with ordinary least squares and semilog specifications. The results are similarly ambiguous.

4.3. *Experiment 3*

To normalize the distribution of family wage and salary income, a Box-Cox transformation (Box and Cox 1964) is tested. The formula for the Box-Cox transformation is as follows

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

The optimal value for lambda (λ), found using a maximum likelihood estimation technique described by Scott and Rope (1993), is 3/8. This value (which is confirmed by a nonlinear regression) is particularly interesting because it is almost exactly half way between 1 (i.e., weighted least squares is appropriate) and 0 (i.e., semilog modeling is appropriate).

The U.S. Bureau of Labor Statistics process of ordinary least squares and stepwise regression is conducted on the transformed values of family wage and salary income, and a reduced model is found. To further confirm that the transformation is appropriate, Experiment 1 is performed on the Box-Cox results. The Box-Cox results outperform the semilog specification in this test (i.e., mean squared error is 460,250,491 for the Box-Cox). The superiority of the Box-Cox specification is also confirmed with the Johnson-McClelland test (forthcoming), a nonparametric specification test designed to find relationships between regressors and disturbance terms. Only under the Box-Cox specification is the null hypothesis of correct specification not rejected.

Further evidence of the superiority of the Box-Cox specification comes from an examination of the residuals of the three models. Only the error terms from the semilog and Box-Cox models appear to have a mean of zero. For the weighted least squares model the mean of the residuals is 1,428.51, with an associated t-statistic of 3.51. Clearly, the hypothesis of a zero mean can be rejected even at the 1% significance level. Additionally, although none of the models produces normally distributed error terms, the Box-Cox model comes the closest, with skewness of 0.41 and kurtosis of 4.69. The semilog model is second best (skewness: -1.39; kurtosis: 5.41), followed by the weighted least squares model (skewness: 7.63; kurtosis: 122.67).

4.4. *Conclusions from the experiments*

Although the first experiment indicates that the semilog specification is superior to the weighted least squares specification, the second experiment yields ambiguous results. In the third experiment a Box-Cox transformation of the income variable is tested and found

to be superior to both the weighted least squares and semilog specifications. Therefore, the Box-Cox specification is used henceforth.

5. Results

Although a final method for imputation has not yet been selected, the predicted values from the Box-Cox model are useful to analyze. Table 1 shows results from this model using only valid salary reporters; i.e., at least one person reports a salary amount, and no one has an invalid blank (such as a refusal to answer) for salary. This reduces the (unweighted) sample 5% to 2,607 consumer units.

The signs for most parameter estimates make sense intuitively. But the signs for the age and education coefficients seem counterintuitive at first. However, when the interactions for age and age squared with education are taken into account, the expected relationships hold in most cases. For example, the parameter estimate is negative for age of the principal earner, and positive for age squared. Similarly, the parameter estimate for education of the principal earner is negative. However, the parameter estimate is positive for the interaction of age and education, and negative for the interaction of age squared and education. When education is held constant at nine or more years, the sum of the parameter estimates for age and its interaction with education is positive, and the sum of the parameter estimates for age squared and its interaction with education is negative. Because most principal earners in the sample have at least completed the ninth grade, the parameter estimates for most of the sample are plausible when interaction terms are taken into account.

Table 2 presents the results of a simple imputation process in which the predicted values from the Box-Cox model are substituted for invalid income reports. (Paulin and Sweet 1995 report additional results.) Although the model is weighted for the population, the unweighted results are shown because the unweighted means are not much different from the weighted means, and weighted standard errors are costly to calculate.

The table shows first the size of the sample for the consumer units under study. For example, of the 3,216 two-member consumer units whose second interview occurs between 1988 and 1990, 2,780 are complete income reporters. Of these, 283 report less than 10,000 USD for total income before taxes. (It should be noted that total income before taxes is not collected directly in the Consumer Expenditure Survey. It is obtained by adding reported values for wages and salaries, self-employment income, Social Security and Railroad Retirement, and supplemental security income for all members to several other sources, such as interest income, collected for the consumer unit as a whole.)

The next tier of information describes wage and salary income for all consumer units. The "Observed" line shows the average wage and salary income that is actually reported for each group. The "Imputed" line shows the average wage and salary income obtained when the predicted values are substituted for the invalid values. Similarly, the "Observed" and "Imputed" lines under "Total Family Income Before Taxes" describe means of actual data reported, and means of income before taxes when reported wage and salary income is replaced by the "imputed" values of the wage and salary income.

The last tier, titled "Predicted Wage and Salary Income for Families with at Least One Invalid Blank," describes mean predicted wage and salary income for each of the groups. For example, of the 283 consumer units classified as complete income reporters with less

Table 1. Results of merged model

Dependent Variable: $(Y^{\lambda} - 1)/\lambda$
where Y = family wage and salary income
and $\lambda = 3/8$

F Value: 86.204
R^2: 0.6495
Adjusted R^2: 0.6419

Independent variables	Parameter estimates	t-Statistics
Intercept	26.328	1.543
Personal characteristics:		
Age of the principal earner	−3.209	−6.119
Age squared (principal earner)	0.032	5.127
Age of the second person	−0.948	−3.156
Age squared (second person)	0.012	3.290
Years of education[1] (principal earner)	−6.551	−7.787
Years of education (second person)	−1.501	−3.406
Interaction terms:		
Age*education (principal earner)	0.378	8.845
Age squared*education (principal earner)	$-3.96*10^{-3}$	−7.820
Age*years of education (second person)	0.099	3.713
Age squared*education (second person)	$-1.20*10^{-3}$	−3.598
Hours per week worked (principal earner)	0.531	10.646
Hours per week worked (second person)	0.237	4.929
Weeks per year worked (principal earner)	0.237	5.098
Weeks per year worked (second person)	0.041	0.830
Respondent is female	−3.096	−2.986
Principal earner is female	−4.868	−3.838
Principal earner earns a wage or salary	27.829	8.535
No one earns mainly wage or salary income[2]	−18.255	−2.176
Both persons earn mainly wage or salary income[2]	13.264	6.063
Principal earner currently has a job	4.559	2.381
Second person currently has a job	−0.979	−0.512
Family type (Husband and wife only)		
Single parent family	−6.914	−2.043
Other family	−3.014	−2.200
Occupational class (Managers and professionals):		
Principal earner:		
Technical/sales	−9.870	−7.198
Precision/production	−5.490	−2.530
Operative or machinist	−11.135	−6.649
Services	−18.250	−9.751
Second person:		
Technical/sales	−3.453	−2.207
Precision/production	−2.096	−0.645
Operative or machinist	−7.264	−3.329
Services	−9.472	−4.160

Table 1 continued

Independent variables	Parameter estimates	*t*-Statistics
Family lives in:		
Northeast region	4.900	3.862
Rural area	−10.313	−6.919
Family reports:		
Owning (not renting) its home	6.169	4.595
Business income (Self-employment or own farm)	−12.559	−7.108
Social Security income	−9.504	−4.765
Pension income	−3.046	−1.827
Interest income	4.903	4.453
Welfare/other government-sponsored income	−0.865	−0.564
Other income	−4.064	−2.480
Individual Retirement Account (at least one)	4.859	3.485
Work status indicators:[3]		
Work status (A)	31.300	9.679
Work status (B)	26.144	9.523
Work status (C)	24.901	8.723
Work status (D)	22.395	10.784
Work status (E)	13.215	7.234
Work status (F)	4.552	1.593
Work status (G)	15.185	3.758
Work status (H)	5.872	2.008
Other characteristics:		
Length of interview (in minutes)	0.025	1.690
Interview occurs in last three months of year	−3.895	−3.357
Number of rooms in consumer unit living quarters	1.075	3.163
Level of the Consumer Price Index[4]	0.519	5.289
Housing unit is public housing	−17.448	−2.244
Government pays other parts of housing costs	−19.366	−3.169

Omitted groups shown in parentheses where appropriate.

[1] 0 years means no schooling; 18 years means at least 2 years of graduate school.

[2] Based on description of occupation. If the job for which the member received the most earnings during the last 12 months is a wage and salary position, that member is defined as earning mainly wage or salary income.

[3] See Appendix C for description of categories.

[4] During month of interview.

than 10,000 USD before taxes, 67 consumer units have at least one invalid blank for wage and salary income. The average predicted wage and salary income for these families is 7,575 USD. By substituting the predicted values of wage and salary income for the reported values of the consumer units with invalid blanks, the average wage and salary income for the 283 increases from 4,005 USD to 5,473 USD. The average income before taxes for the 283 increases from 5,866 USD to 7,334 USD.

Finally, there are lines for each group indicating "minimum" and "maximum" values reported for each group. Because the complete reporters are separated in the table by total income before taxes, note that it is possible for the maximum wage and salary income reported to exceed the level of income before taxes reported. For example, the maximum

Table 2. Effects of replacing invalid blanks with predicted wage and salary data for two-member consumer units, 1988–1990

Variable	All consumer units	Complete income reporters	Income before taxes: Complete income reporters only					Incomplete income reporters
			Under $10,000	$10,000 to $19,999	$20,000 to $29,999	$30,000 to $39,999	$40,000 and Over	
Sample size	3,216	2,780	283	507	541	445	1,004	346
Family wage and salary income:								
Observed	**$28,519**	**$31,505**	**$4,005**	**$10,868**	**$20,070**	**$29,692**	**$56,642**	**$4,532**
Std. Err	529	539	236	255	329	391	1,008	1,496
Minimum	0	0	0	0	0	500	0	0
Maximum	520,000	520,000	44,000	23,157	43,200	44,000	520,000	500,000
Imputed	**$30,366**	**$31,920**	**$5,473**	**$11,610**	**20,489**	**$29,936**	**$56,670**	**$17,882**
Std. Err	513	536	354	296	356	410	1,007	1,571
Minimum	0	0	0	27	11	500	351	0
Maximum	520,000	520,000	44,000	47,593	58,339	58,150	520,000	500,000
Total family income before taxes:								
Observed:	**$34,075**	**$37,504**	**$5,866**	**$14,997**	**$24,862**	**$34,375**	**$65,986**	**$6,528**
Std. Err	607	593	230	134	123	136	1,115	2,222
Minimum	−25,920	−25,920	−25,920	10,000	20,000	30,000	40,000	0
Maximum	750,000	524,000	9,983	19,970	29,976	39,943	524,000	750,000
Imputed	**$35,923**	**$37,919**	**$7,334**	**$15,740**	**$25,281**	**$34,619**	**$66,014**	**$19,879**
Std. Err	590	591	356	205	179	171	1,121	2,249
Minimum	−25,920	−25,920	−25,920	10,000	14,028	25,558	34,185	33
Maximum	750,000	524,000	52,457	53,004	58,489	59,950	524,000	750,000
Predicted wage and salary income for families with at least one invalid blank:								
Predicted	**$16,868**	**$17,502**	**$7,575**	**$14,693**	**$27,949**	**$41,444**	**$46,614**	**$16,547**
Std. Err	686	1,393	1,183	2,120	2,805	3,980	6,865	757
N	470	158	67	44	28	11	8	312
Minimum	0	0	0	27	11	22,173	18,132	0
Maximum	72,459	72,459	43,457	47,593	58,339	58,150	72,459	71,400

wage and salary income reported for families with less than 10,000 USD in total income before taxes is 44,000 USD. This result is possible because this family presumably had at least 34,001 USD in business or other losses that offset the wage and salary income.

Table 2 shows that even complete income reporters have higher average incomes when salary is estimated from the model. Although the difference is small, the fact that *any* complete reporters need to have salary imputed confirms that the complete income reporter definition does not fully correct for income reporting problems. Differences will probably be greater when estimates for other sources are also included. Even so, in each group with less than 40,000 USD in income there is at least one consumer unit predicted to have more than 50,000 USD in total income when salary *alone* is estimated from the model, indicating that imputation is useful and necessary.

6. Conclusions and Future Work

In this study wage and salary income data from the 1988–90 U.S. Consumer Expenditure Survey are examined. A large portion (15%) of the sampled families are classified as incomplete income reporters, and not all complete income reporters provide a full accounting of all sources of income. To reduce the problems of income nonresponse, model-based imputation is explored as a strategy for wage and salary income.

The data are assumed to be missing at random – that is, the propensity to respond to income questions is related to demographic characteristics, but not directly to level of income. Two-member consumer units are analyzed because they represent a link between single-member consumer units (which have few inherent difficulties for modeling) and more complex multiple-member consumer units (which have many inherent difficulties).

Although each agency has similar goals, different modeling procedures are examined by the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census. For example, in order to minimize heteroscedasticity, the U.S. Bureau of Labor Statistics uses a weighted least squares model; the U.S. Bureau of the Census uses a semilog model. As another example, each agency attempts to find variables with maximum explanatory value to minimize production costs. However, the U.S. Bureau of Labor Statistics uses a stepwise method; the U.S. Bureau of the Census develops a method of "Transformed Main Effects" to analyze main effects and interaction terms.

When each agency finds the best model using its own strategies, variables from each are merged into a suggested model for use in imputation, and the stepwise procedure is used again to select the most predictive variables. Because different specifications (weighted least squares and semilog) are used, which specification to use for the final imputation model is an issue to be resolved. Using a Box-Cox transformation, an intermediate specification is proposed. After a series of tests and examination of each model's residuals, evidence indicates that the Box-Cox transformation produces the best model.

Finally, preliminary estimates suggest that imputation will affect published data. When predicted data for consumer units with at least one invalid wage and salary report are substituted for the wage and salary income actually reported for the consumer unit, average wage and salary income increases 415 USD for complete income reporters, and

1,848 USD for all consumer units (that is, complete and incomplete reporters). Although the final imputation process includes the addition of random noise to predicted values to properly preserve the variance of the income distribution, the increase in means indicates that the model results move the data values in the proper direction.

More work must be done before final imputation models can be recommended. The next step is to apply the lessons learned in this study to multiple-member consumer units. Other income sources must also be analyzed. Additionally, the missing at random assumption needs more investigation, and there are experiments underway to test this assumption. However, missing at random assumptions have yielded new models for examination, as described in this article. These results provide a valuable foundation for further research.

Appendix A. Consumer Unit Definition

Consumer units (the basic unit of comparison in the Consumer Expenditure Survey) are defined as a single person either living alone or sharing a household with others from whom the single person is financially independent; two or more members of a household related by blood, marriage, adoption, or other legal arrangement; or two or more persons living together who share responsibility for at least two out of three major types of expenses – food, housing, and other expenses. For convenience, ''family'' and ''consumer unit'' are used interchangeably in the text.

Appendix B. Statistical Tests

This section provides a brief description of several of the statistical tests mentioned in this article.

B.1. The Chow test

The Chow test is designed to test whether or not two samples can be pooled for use in regression. First, regressions with identical dependent and independent variables are run separately for each group of interest (in this case, husband/wife and single parent/other families). Then the samples are pooled, and the identical regression is run. Using the results, an F-statistic is computed. If it is statistically significant, the samples should not be pooled.

B.2. The Breusch-Pagan test

In the Breusch-Pagan test the residuals from the initial wage and salary regression are squared and regressed on independent variables suspected of being related to the error terms. A ratio symbolized by λ is created. The numerator is the regression sum of squares from the regression of the squared residuals on independent variables. The denominator is a number twice as large as the square of the mean squared error from the initial wage and salary regression. The ratio λ has a χ^2-distribution with degrees of freedom r, where r is the number of independent variables in the equation where the squared residuals are regressed on independent variables. See Maddala (1988), pages 164–167 (especially page 165) for more information.

B.3. The Goldfeld-Quandt test

In the Goldfeld-Quandt test the data are arranged in order from lowest to highest value of the independent variable assumed to be associated with heteroscedasticity. Because predicted wage and salary income is the only independent variable in these models, the data are ordered with respect to this variable. Some portion of the "middle" observations are omitted, and two separate regressions are run on the set of data containing the lowest values for the independent variable (set 1) and the highest values for the independent variable (set 2). The error sum of squares from the set 2 regression is divided by the error sum of squares from the set 1 regression. An F-test is used to determine whether this ratio implies that the error sums of squares are different for sets 1 and 2. If so, this is evidence of heteroscedasticity.

B.4. The Park-Glejser test

In the Park-Glejser test the true error variance is assumed to be a multiplicative function of the independent variables in the regression equation, each raised to some power δ. In order to estimate δ, the natural logarithm of the error variance is regressed on an intercept and the natural logarithm of each independent variable. Because there is only one independent variable (predicted value of wage and salary income) in the model under study, the model is specified as follows

$$\log\varepsilon_i^2 = \gamma + \delta \log X_i + u_i$$

If δ is statistically significant, this is evidence of heteroscedasticity.

B.5. The Bera-McAleer test

In the first stage of the Bera-McAleer test two regressions are run, each with identical independent variables. However, in the first regression the natural log of wage and salary income (in the present case) serves as the dependent variable. In the second regression observed wage and salary income serves as the dependent variable. The second stage then uses the predicted values from these first two regressions. First, predicted natural log of income is exponentiated, and the resulting variable is regressed on the original independent variables. Using Maddala's notation, the error term, v_{1t}, is calculated for use later. Similarly, the natural log of predicted income is regressed on the original independent variables, and the error term, v_{0t}, is also calculated. In the third stage the original regressions are rerun, except that v_{1t} is now used as an independent variable in the semilog model, and v_{0t} is used in the linear model. If the parameter estimate is statistically significant for v_{1t}, but not for v_{0t}, the linear specification is better. If the parameter estimate is statistically significant for v_{0t}, but not for v_{1t}, the semilog specification is better. If both parameter estimates are, or if neither is, statistically significant, the results are inconclusive.

B.6. The PE test

In the PE test the first stage is identical to the Bera-McAleer test. However, variables for the second (and final) stage are different. First, the predicted natural log of income is

exponentiated and subtracted from the predicted value of income. The value becomes a variable denoted here as "Y_0" for convenience. Second, the natural log of predicted income is subtracted from the predicted natural log of income. The value becomes a variable denoted here as "Y_1." In the second stage the natural log of income is regressed on the original independent variables and Y_0. In a separate regression observed income is regressed on the original independent variables and Y_1. The parameter estimates for Y_0 and Y_1 are compared. The same rules of statistical significance used in the Bera-McAleer test also apply in the PE test.

Appendix C. Job Status Indicator Variables

Categorical variables

Work status

A (very good jobs for both persons)

Consumer units where both work full time/full year (ft/fy) and both employers contribute to the pensions.

B (good jobs)

Consumer units where either both people work ft/fy and one of the persons receives the pension

or

consumer units where one person works ft/fy and the other person works but not ft/fy but both persons' employers contribute to their pensions (i.e., good part time job).

C (employed persons, but with no retirement benefits)

Consumer units where both work ft/fy but neither employer contributes to their pensions.

D (single earner with good job, or two earners with different status)

Consumer units where only one person works ft/fy, and only one person's employer contributes to a pension,

or

consumer units where neither person works ft/fy but both persons' employers contribute to pensions (i.e., so both persons are working).

E (consumer units do not have as high paying salaries as other consumer units)

Either there is only one person working ft/fy with no employer contribution to pensions

or

consumer units with no persons working ft/fy but there is one employer contributing to a pension.

F (poor salaries but working)

Consumer units where no one works ft/fy and no one's employer contributes to a pension, yet both persons in the consumer unit currently have a job.

G (odd cases)

Consumer units where both members work ft/fy yet at least one valid blank for whether or not the employer contributes to the pension.

H (invalid blank)

All consumer units where there was a G response for the employment contribution variable.

Omitted group:

I (poor salaries and not working)

Consumer units where no one works ft/fy, neither persons' employer contributes to a pension, and at least one person in the consumer unit currently does not have a job.

7. References

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society, No. 2, Ser. B, 211–243, especially p. 214.

Crawford, S. (1989–90). Internal Memoranda. U.S. Bureau of Labor Statistics, dated 1989 through 1990.

Crawford, S. (1989). Internal Memorandum. U.S. Bureau of Labor Statistics, March 3, 1989, p. 6.

Greenlees, J.S., Reece, W.S., and Zieschang, K.D. (1982). Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed. Journal of the American Statistical Association, 77, 251–261.

Johnson, D.S. and McClelland, R. (forthcoming). Nonparametric Tests for the Independence of Regression and Disturbances. Review of Economics and Statistics. Unpublished version available as U.S. Bureau of Labor Statistics Working Paper 231, Office of Prices and Living Conditions, June 1992.

Kennedy, P. (1992). A Guide to Econometrics (3rd ed.). Cambridge, MA: The MIT Press, 108–109.

Little, R.J.A. (1993). Discussion. Proceedings of the Section on Survey Research Methods, American Statistical Association, 117–118.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley & Sons.

Maddala, G.S. (1988). Introduction to Econometrics. New York: Macmillan Publishing, 130–137 and 159–180.

Paulin, G.D. and Ferraro, D.L. (1994). Imputing Income in the Consumer Expenditure Survey. Monthly Labor Review, December, 23–31.

Paulin, G.D. and Sweet, E.M. (1995). Modeling Income in the U.S. Consumer Expenditure Survey. U.S. Bureau of Labor Statistics Statistical Note Series, No. 37, June.

Pindyck, R.S. and Rubinfeld, D.L. (1981). Econometric Models and Economic Forecasts (2nd ed.). New York: McGraw-Hill Books, 150–152.

Scott, S. and Rope, D.J. (1993). Distributions and Transformations for Family Expenditures. Proceedings of the Section on Social Statistics, American Statistical Association, 741–746.

Received February 1994
Revised September 1995