Journal of Economic and Social Measurement 22 (1996)
103-128 IOS Press

Do expenditures explain income? A study of variables for income imputation

Geoffrey D. Paulin
U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE 113985, Washington, D.C, 20212. USA
Phone:(203) 606-6900; Fax: (202) 606-7006

David L. Ferraro *US. Bureau of Census. Room 3784-3*. Washington, *D.C. 20233. USA*

Income data in the U.S Consumer Expenditure Survey are subject to nonresponse. Model-based imputation is being explored to diminish missing data problems. Since income is an important variable to predicting expenditures, might expenditures be useful in predicting incomes? Incomes from wages and salaries and self-employment are modeled. These results are compared to regressions on demographic characteristics alone. Although each expenditure category adds to the predictive power of the model, total expenditures adds the most.

1. Introduction

Income is one of the most important variables in studies of consumer spending patterns. Not only is it an important predictor of expenditures, it has also been used to predict the probability of events as diverse as the consumption of wine [2] and the purchase of a home [4, 8]. Yet, nonresponse to income questions is a common problem in household surveys. The U.S. Consumer Expenditure Survey (CE) is no exception.

Many studies have been conducted to ascertain the best way to fill in the blanks For example, Lillard et al. [12] and David et al. [6] examine methods for imputing income in the Current Population Survey. Eltinge and Yansaneh [7] pursue weighting adjustment as a method to estimate mean consumer income in the CE. Paulin and Sweet [16] experiment with model building to estimate wage and salary income for individual nonrespondents in the CE. None of these studies directly uses expenditure data to impute income.

Given the problems associated with using a partial data set, and the fact that income is so often used to predict expenditures, two questions come to mind: Do expenditures explain income? If so, should they be used that way? This study examines the predictive power of similar equations, one without expenditures and the others including different expenditure variables. These equations are used to predict labor income (wage and salary, self-employment income) for single persons. The predictive powers of these equations are compared using various statistical techniques.

0747-966J96/58.00 ©1996 - IOS Press .411 rights reserved

2. The survey

Sponsored by the U.S. Bureau of Labor Statistics (BLS) and collected under contract by the U.S. Bureau of the Census, the CE is the only major U.S. government survey to collect detailed expenditure and demographic data from families. The information is collected from participating consumer units[1] in a series of five quarterly interviews. Income data are collected in the second and fifth of these interviews.

Currently, consumer units are divided into two groups: "complete" and "incomplete" income reporters, depending on the respondent's answers to income questions[2]. 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. As a result, these classifications do not completely correct for the problems caused by missing data. For example, many groups are shown on average to spend more than their reported incomes, even though only complete reporters are used to define income classes. It is hoped that imputing data to replace missing income values will improve the quality of the published CE data.

3. Prediction issues

3.1. Data issues

Based on the terminology of Little and Rubin [13], the data are assumed to be missing at random (MAR). If the data are MAR, the propensity to respond may be related to any number of demographic characteristics (such as age, number of earners, and other factors), but it is not related directly to level of income. Support for the MAR assumption is described in other recent work using CE data [15, 16].

3.2. The sample

Because there are questions to be addressed about whether expenditures should be used at all in the imputation process, preliminary investigation should be as simple as possible, consistent with achieving useful results. For this reason only single-member consumer units are studied. In this way the earner and the decision maker are the same person, and there are no questions about infra-household

[1] Consumer units (the basic unit of comparison in the CE) 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 2 out of 3 major types of expenses - food, housing, and other expenses.

[2] For more information about sources of income collected and the definitions of complete and incomplete reporters, see Paulin and Ferraro [15, pp. 30-31).

resource sharing or other interactions among persons for which to account when incorporating expenditures into the models. Further, singles comprise a large enough portion of the sample (29% of consumer units interviewed in 1992) to constitute a sufficiently important group to study.

The sample is further restricted to consumer units in their second interview who said that most of their earnings during the past year had come from work at a particular occupation, either self-employed or working for wage and salary income, as opposed to pension or other supplemental income. Those who report that most of their earnings are from a wage and salary occupation are defined as salaried singles, even if they have some self-employment or other income; similarly, those who report that most of their earnings are from self-employment are classified as self-employed singles, even if they report wage and salary or other income. Additionally, only "valid" reporters of income are included. Salaried singles are included as valid reporters if the respondent reports positive wage and salary income. Self-employed singles are included if the respondent reports no negative self-employment income either from business or farm, and if neither source of income (business or farm) has an invalid response (refusal or "do not know").

The salaried singles were interviewed between 1988 and 1990. (These years are chosen because they are the internal data to which Census had access under an interagency agreement.) The self-employed data are from interviews taking place between 1988 through 1992 in order to achieve a large enough sample for study. In theory. the sample could be broadened by including single-earner families, and adding independent variables to the models to control for factors such as number and age of children, marital status of the earner, and interactions between expenditures and family size, age of oldest child, and other characteristics. However, Chow tests [10] show that single-earner families and single persons are distinct groups, regardless of type of income earned.

The total sample size for the salaried singles is 2,247. The total sample size for the self-employed singles is 207. In the regression results described below, a small number of consumer units have a missing value for one of the independent variables (length of interview). Therefore, the sample size for the regressions is 2,207 for single salaried consumer units, and 202 for the self-employed.

3.3. Selecting expenditure categories

Choosing appropriate expenditure variables presents a challenge. Those with an identifiable Engel curve (i.e., expenditures as a function of income) are the best candidates. The stronger the relationship, the more obvious the shape of the Engel curve should be, and the more useful the expenditure information becomes in predicting income. (The shape can be postulated from scatter plots; the strength of the relationship can be gauged by examining t-statistics or performing

standard specification tests.) Perhaps the most obvious candidate is total expenditures, since these data should clearly be related to major sources of income. However, other expenditures may also be significantly correlated with income. It is even possible that some subcategories of expenditures may be better predictors than total expenditures. For example, the Engel curve for a specific expenditure may be estimated with a lower variance than the Engel curve for total expenditures.

But many expenditure categories have disadvantages that total expenditures do not. For example, virtually all consumer units have some value reported for total expenditures, but not every consumer unit incurs every type of expenditure. Therefore, it is important that if specific expenditures are chosen, there should be few non-purchasers; or, if there are a substantial number of zeros reported, it is important that those zeros be meaningful in the present context. That is, if almost no one under a certain income ever purchases a certain item, and almost everyone with more than the critical amount makes a purchase, then the zero expenditure may yield useful information. But if purchases of the item are naturally lumpy over time regardless of income (e.g., automobile purchases), then the lack of an expenditure is not a meaningful indicator of level of income.

Of further interest is whether the income elasticity of a particular expenditure (i.e., the percent change in the expenditure due to a one percent increase in income) might play a role in predicting income. For example, items with a low income elasticity (i.e., less than one) may help predict wage and salary income, being that these incomes are relatively stable, whereas items with a high elasticity (i.e., greater than one) may better predict the more transitory self-employment incomes, since high elasticity items are by definition more sensitive to changes in income. Based on Logit results (to predict the probability of reporting the expenditure given income and other characteristics), graphical analysis (i.e., plots of level of expenditure on income), and statistical comparisons, three candidates are selected. These are: food at home, shelter and utilities, and telephone services. However, the endogeneity issue becomes more complicated if individual expenditure categories are used for imputation instead of total expenditures. For example, some endogeneity may exist for all data users if total expenditures are used. But if food at home is chosen for use in imputation, researchers analyzing housing demand will have little concern with endogeneity whereas researchers analyzing food demand will have a greater concern. To help address these issues, some compromise candidates are proposed, based on Paulin [14]. In this paper, the author examines the relationship between housing tenure choice (i.e., whether to own or tent one's home) and expenditures. Two categories he studies are basic goods and services (i.e., the sum of food at home, shelter and utilities,, and apparel) and recreation and related expenditures (i.e., the sum of entertainment, food away from home, vacation and other housing, and reading). These groups are chosen as candidates here because they each have few zeros, the endogeneity

problem is lessened with summed expenditures, and *a priori*, they should have different income elasticities (basic goods should be low and recreation should be high). Reading is not included in the present definition of recreation and related expenses because its value is too small and infrequently reported to be useful in predicting incomes.

In each case, expenditures are deflated by the level of the Consumer Price Index (CPI) for all goods and services for the month in which the interview takes place. This puts expenditures in real dollar terms, because the CPI measures the rate of changes in prices. It is important to control for price changes in some way because multiple years of data are used in each sample, and because expenditures, if used ultimately in imputation, will change as prices do; the division by CPI puts expenditures in real dollar terms. Incomes are also deflated by the CPI before being included in the regressions; thus, real dollar incomes are predicted from the models using real dollar expenditures.

3.4. Demographic variables

Other independent variables are also included in the models to predict income, as outlined in Appendix B. These include demographic variables identifying age, education (including a dummy variable for current student status), race, and sex of the respondent; dummy variables for location (region of the country and urban/rural area) and tenure (owner or renter) of housing.

In addition to the continuous demographic variables (age, age squared, and education), interaction terms are included (age*education and age squared*education). These interactions, which are found to be important in Paulin and Sweet [16], also have significant explanatory power in the present models, at least for the salaried singles (Table 8). It should also be pointed out that although the education variable is continuous, and reflects the number of years of education of the respondent, the maximum value that is available for this variable is 18, or at least two years of graduate school.

Housing tenure is included because results of the Interview survey routinely show that homeowners report higher incomes than renters. Additionally, a dummy variable is included describing whether the person, if a homeowner, still owes for a mortgage. The interaction of this dummy variable and the level of the expenditure variable is used in each regression. Paulin [14] finds that owners with and without mortgages differ frequently in expenditure pattern, even when income and other characteristics are controlled, probably because a mortgage, once negotiated, is a fixed cost for the consumer. A person with the same level of income could choose to save the amount that would have gone toward a mortgage, or spend it on goods and services other than housing. The interaction term helps to account for these differences.

3.5. Labor-related variables

Attributes of the respondent's occupation (dummy variables describing type of occupation, whether other forms of labor income are also earned, and other variables describing number of hours per year worked) are included in each model. Perhaps the most interesting of the independent variables are those describing hours per year worked. When viewing scatter plots of labor income by hours per year worked (HOURYEAR), changes in the relationship are observed at different points. Although this is particularly we of self-employment income, in each case there is a spike at 2,080 hours (40 hours per week, 52 weeks per year), and a noticeable slope change for those working more than 2,080 hours per year. For self-employed persons the slope appears to be close to zero. To account for these discontinuities, dummy variables and interactive terms are used. For those who work exactly 2,080 hours, the variable FULLTIME equals one; for all others, it equals zero. For those who work more than 2,080 hours, the variable OVERTIME equals one; for all others, it equals zero. Finally, OVERTIME is interacted with HOURYEAR to form OTSLOPE (i.e., OTSLOPE = OVERTIME * HOURYEAR). This allows the change in slope for those who work overtime to be measured.

Somewhat related to these issues is that of whether the respondent works two jobs, as may be the case when the respondent reports more than one source of labor income (i.e., wage and salary income *and* self-employment income). David et al. [6] note that their model predicts wage and salary income better for individuals for whom wages and salaries are the only source of income than for those who also report self-employment income, because hours per year is the total worked for *both* sources of income [6, p. 32]. However, it may not be the case that the person *concurrently* works for both sources, but may have changed jobs at some point in the last year. In either event, there may be other hidden relationships between the income sources for a person who has both. For example, a person whose main income is self-employment may earn supplemental salaried income if business is slow. Or a salaried earner may choose to work fewer hours away from home if some self-employment work at home is available. To make some attempt to control for these complicated possibilities, a dummy variable is included in each regression if the earner reports receipt of both types of labor income.

3.6. Survey attribute variables

Variables describing survey attributes including length of interview (continuous); quarter of the year in which the interview took place (binary); and for the self-employed, whether the interview occurred after 1990 (binary) are included for several reasons. Persons with longer interviews may have more expenditures or income information to report than those with shorter interviews. Incomes may

also be better reported during certain quarters of the year due to their proximity to the tax season. For the self-employed a dummy variable RECESS is included for those who are interviewed in 1991 or 1992. This variable serves a dual purpose. According to *Survey of Current Business* [18], Gross Domestic Product, when measured in constant 1987 dollars, experienced two consecutive quarters of negative growth in the last quarter of 1990 and the first quarter of 1991 [18, p. 3]. Since the respondent is asked to recall income for the previous year, each respondent interviewed between January 1991 and December 1992 is asked to recall. income for which at least one full quarter falls during the period of negative economic growth. At the same time, this variable controls for the difference in sample periods between the self-employed and salaried singles, since salaried singles are included for 1988-90 only. Table 9 shows that the coefficient for RECESS is negative but not statistically significant, suggesting the weakening economy had a small negative effect on average self-employment income, but this result is not conclusive.

3.7. Transformations

Neither income data nor expenditure data are often found to have a normal distribution. Many authors [6, 9] use the log of income in their models to approximate normality. Although log transformations have some desirable properties, it is not clear that the log transformation is optimal for approximating normality, nor that the best results are obtained by predicting log of income. Paulin and Sweet [16] find that the best results are obtained using a more general transformation is described by Box and Cox [3]. The formula for such a transformation is:

$$(Y^{\lambda} - 1)/\lambda,$$

where Y is the variable being transformed and λ is a parameter. An important feature of the Box–Cox transformation is that if λ equals one, no transformation is necessary; however, if λ approaches zero, the log transformation is appropriate. Furthermore, λ can take on any value; thus, the approximation to normality can be closer than when the log is chosen arbitrarily, and, as with the log transformation, heteroscedasticity is substantially reduced. The optimal value of λ is found through a maximum likelihood estimation procedure described by Scott and Rope [17].

In the regressions described below, the optimal value for λ for income is 0.375 for single salaried workers and 0.200 for single self-employed workers. The value for salaried workers is particularly interesting, because it matches the value that Paulin and Sweet [16] find for wage and salary income for two-member consumer units[3]. Also of interest is that for each type of worker the optimal value of λ

[3] Paulin and Sweet do not divide income by the CPI before transforming, but this division is not expected to have a large effect on the normality of the distribution because the CPI changes slowly and steadily over the period under study.

Table 1
Optimal values of λ for income and expenditures

Variable	Salaried	Self-employed
Income	0.375	0.200
Total expenditures	0.125	0.125
Food at home	0.425	0.375
Shelter/utilities	0.475	0.375
Telephone services	0.375	0.375
Basic goods/services	0.350	0.250
Recreation related	0.200	0.275

for total expenditures (0.125) and telephone expenditures (0.375) match, indicating a similar distribution of these expenditures regardless of source of income. The values of λ are summarized in the Table 1.

For the remainder of this paper, when income or expenditures are discussed, it is the transformed variable that is being described, unless otherwise stated.

3.8. Weighting

The regressions are weighted to reflect the population and to account for sample design effect.

3.9. Multicollinearity

Usually, when the goal is to impute a variable, multicollinearity in the model stage is not a serious problem. The reason is that it is the predicted outcome, and not any individual parameter estimate, that is of interest. However, in the present case, it is important to know whether expenditures are highly collinear with demographic characteristics for implementation. If expenditures are perfectly explained by the other independent variables, then it is more efficient to include only expenditures in the model. On the other hand, if processing is more complicated when expenditures are used, it may be more efficient to use only demographics in the model. Kennedy [10, p. 181] suggests that if the $R2$ from the regression of income on expenditures and other demographic characteristics exceeds the $R2$ for the regression of expenditures on the other demographic characteristics, multicollinearity is not a serious problem. When expenditures are regressed on characteristics, the largest $R2$ is 0.5204 (for total expenditures for the self-employed). Since this value is smaller than the smallest $R2$ for income regressed on an expenditure and other characteristics (0.6167 for self-employed food at home model), multicollinearity is not serious. Table 2 summarizes results of the regressions of expenditures on demographic characteristics.

Table 2
R^2 values for expenditures regressed on demographic characteristics

Variable	Salaried	Self-employed
Total expenditures	0.4316	0.5204
Food at home	0.1425	0.2237
Shelter/utilities	0.3739	0.4854
Telephone services	0.1916	0.3116
Basic goods services	0.3818	0.4499
Recreation related	0.2046	0.3498

Table 3
R^2 values regressions of income on expenditures and other characteristics

Variable	Salaried	Self-employed
No expenditures	0.6498	0.6042
Total expenditures	0.7091	0.7070
Food at home	0.6512	0.6167
Shelter/utilities	0.6690	0.6422
Telephone services	0.6537	0.6209
Basic goods/services	0.6763	0.6617
Recreation related	0.6687	0.6248

4. Results.

4.1. Predictive power of expenditures

A simple test of the power of each expenditure is to compare the R^2 values for each regression to find the largest value. For the salaried singles, when income is regressed solely on demographic characteristics (i.e., expenditures are excluded from the model) the resulting R^2 is 0.6498. When food at home is added, the value increases slightly to 0.6512. When total expenditures are added, the value increases to 0.7091. Similar results are obtained from the self-employed singles. Without expenditures, the R^2 is 0.6042. When food at home is added to the model, the value increases to 0.6167. When total expenditures are added, the value increases to 0.7070. The R^2 values for all models are shown in Table 3.

It is interesting that whether wage and salary or self-employment income is examined, the order of increase in R^2 for each expenditure is the same. That is, in each case food at home adds the least to R^2, which becomes succeedingly larger in the regressions for telephone services, recreation and related expenditures, shelter and utilities, basic goods and services, and finally total expenditures.

Table 4
Comparison of mean square errors (in millions)

Variable	Salaried	Self-employed
No expenditures	165.37	345.56
Total expenditures	133.37	284.90
Food at home	164.76	344.15
Shelter/utilities	155.65	332.00
Telephone services	163.73	339.23
Basic goods/services	152.49	323.41
Recreation related	155.45	331.12

9.2. Mean square error comparisons for actual income

The models described above are designed to predict transformed incomes. But the real goal of imputation is to predict actual income values. How well do expenditures predict actual income? In order to answer this question, a comparison of mean square errors (MSEs) for each expenditure category is proposed. In this test, the transformed value of income is untransformed in the following way:

$$Y' = (\lambda y' + 1)^{1/\lambda},$$

where y' is the predicted value of transformed income, λ is equal to 0.375 for wage and salary income and 0.2 for self-employment income, Y' is the predicted value of actual income.

The MSE is then found by the formula:

$$\text{MSE} = \sum (Y - Y')^2 / n,$$

where Y is observed income, n is the number of observations of Y'.

Although "untransforming" the dependent variable in this way in theory can cause bias in the error term [16], in, practice the bias is found to be of little consequence, as shown later (Table 7). Table 4 summarizes the MSEs for the equations with and without expenditures.

Once again, the results of the MSE test are similar for both salaried and self-employed singles. The largest MSE (and therefore the least tight fit) is found for the model in which no expenditures are included. The variables, in descending order, of MSE are: food at home, telephone services, recreation and related expenditures, shelter and utilities, basic goods and services, and total expenditures; the same order is seen when R2 values for the models using transformed variables are compared (Table 3).

Additionally, the Wilcoxon rank-sum test [I l] is a distribution-free test used to ascertain whether or not the sum of squared errors (SSE) of the model with no expenditures is equal to the SSEs of the various models with expenditures. For

Table 5
Wilcoxon rank-sum test comparing expenditure models to no expenditure model

	Salaried		Self-employed	
Variable	Z	Prob > |Z|	Z	Prob > |Z|
Total expenditures	−2.7299	0.0063	−1.7217	0.0851
Food at home	−0.0753	0.9400	0.1631	0.8704
Shelter/utilities	0.6972	0.4857	−0.1828	0.8550
Telephone services	0.3324	0.7396	−0.0549	0.9558
Basic goods/services	1.2459	0.2128	−0.5918	0.5540
Recreation related	0.7327	0.4637	0.3464	0.7290

both the salaried and self-employed models, only the total expenditures models are found to have an SSE that is different in a statistically significant way from the model using no expenditures. Although the test results are stronger for the salaried singles (significant at the 1% confidence level) than for the self-employed singles (significant at the 10% confidence level), none of the other models tested display differences that are significant at any level even close to an accepted level of statistical confidence, as shown in Table 5. (The Z-variable is the standardized W-test statistic.)

4.3. Comparisons of means and standard errors

Another way to determine which models are most useful is to compare the means and standard errors of the predicted incomes to the actual incomes to see which models produce the closest results. (For accuracy in comparisons, consumer units for which no length of interview is recorded are omitted before the mean and standard error of the actual income values are calculated, since these observations are also omitted from the regressions.) The income data shown in Table 6 are for the untransformed values. The standard errors of each mean are shown in parentheses below the mean. All statistics in Table 6 are unweighted.

Table 6 indicates that the standard errors of the means of values predicted from the models are significantly lower than the observed values. However, Little and Rubin [13, p. 61] suggest that before the imputation process is considered complete, a random residual value should be added to the value predicted by the regression model to reflect the uncertainty in the predicted value. Before calculating the means shown in Table 7, residuals are added to each predicted observation using the following formula:

$$Y'' = Y' + (\mathrm{MSE})^{1/2} * Z,$$

where Y' is the predicted value of transformed income for a specific observation, MSE is the mean square error from the regression model, Z is an independent standard normal random variable.

Table 6
Means and standard errors of income: observed and predicted

Variable	Salaried	Self-employed
Observed	$15,953	14,830
	(346.17)	(1,468)
No expenditures	14,428	11,034
	(193.28)	(612.86)
Total expenditures	14,642	11,996
	(206.33)	(765.08)
Food at home	14,434	11,116
	(193.29)	(636.62)
Shelter/utilities	14,492	11,369
	(198.52)	(654.53)
Telephone services	14,431	11,128
	(193.44)	(627.11)
Basic goods/services	14,521	11,490
	(200.24)	(685.91)
Recreation related	14,502	11,215
	(197.32)	(650.82)

Table 7
Comparison of observed and predicted means when random noise is added

Variable	Salaried	Self-employed
Observed	$15,953	14,830
No expenditures	15,753	15,105
Total expenditures	15,701	14,707
Food at home	15,748	15,179
Shelter/utilities	15,743	15,145
Telephone services	15,747	15,117
Basic goods/services	15,737	15,077
Recreation related	15,749	15,026

This new predicted value is then untransformed as described earlier. Using the overall MSE in this way is simplistic; however, use of a more appropriate method [5] is beyond the scope of this paper.

An even more striking anomoly with the distribution of the transformed predicted values (Table 6) is that the mean of the predicted values is much lower than the mean of the actual values. This bias in the mean of the transformed values is largely eliminated when the residuals are added to the predicted values before they are untransformed (Table 7). Because all predicted means are well within the 95% confidence interval associated with the observed means, no further statistical testing is performed.

5. Summary, future work, conclusions

Although some work has been done to adjust for nonresponse to income questions in the U.S. Consumer Expenditure Survey, no other work has been completed in which expenditures are used to impute income. Because expenditure data are expected to be correlated with income, and because these data are unique to the CE, this study examines which expenditures yield the most benefit in predicting incomes.

Single persons who have earned most of their income in the last year, either from wages and salaries or self-employment, are studied. Several expenditure categories are also chosen for analysis. Incomes and expenditures are normalized using Box-Cox transformations, which also correct for heteroscedasticity. Transformed expenditures are regressed on other demographic characteristics before being included in the income prediction model to ascertain whether multicollinearity is a substantive problem. Results of a model using only demographic characteristics to predict incomes are compared to results of models that also include expenditures. For each type of income expenditures add to the explanatory power of the regression as measured by R^2. Food at home expenditures add the least to predictive power; total expenditures add the most.

If expenditures are to be used to impute income, total expenditures emerge as the best choice in every method tested here. For researchers using CE data the problem of endogeneity is less for total expenditures than for more specific expenditure categories. Total expenditures also add the most (about 7% for the salaried and 10% for the self-employed) to the R^2 value of these regressions. Finally, regressions using total expenditures have the lowest mean square error, as proven with the Wilcoxon test.

However, this study only addresses single persons. The relationship of expenditures to income becomes more complex as family size, and particularly number of earners, increases. These relationships warrant fuller examination before expenditures can be recommended for use in imputation.

Appendix A: A preliminary experiment with income shares

According to Bannock et al. [1], Engel's original proposition of 1857 is that as incomes increase, the proportion of income spent on food diminishes [1, p. 140J. Because shares of other goods and services may also vary with level of income, it is worthwhile to test some relationships. However, because income is endogenous, it is necessary to use total expenditures as a proxy in the prediction of the share. Thus, the dependent variable in the model becomes the untransformed level of the specific expenditure (i.e., total expenditure, food at home, etc.) divided by the untransformed income from the appropriate source (wage and salary or self-employment income), or the *income shares*. The independent variables include the demographic characteristics and transformed total expenditures.

Surprisingly, none of the shares tested are very useful in predicting income. For the salaried singles the models all have extremely low R^2 values - 0.02 or less in each case. The models also predict negative shares for more than one-fourth of the sample regardless of the model. The coefficient on total expenditures is not statistically significant in any of the models tested; therefore, it is not surprising that the results change little when total expenditures is removed to compare models using only demographic characteristics to predict shares. Part of the problem is that so many respondents report extremely large income shares. For example, the average value of total expenditures divided by wage and salary income is 4.77, meaning that total expenditures are 477% of wage and salary income on average. The most extreme observation is greater than 1200 (or 120,000%), but there are several observations exceeding even 100 (10,000%), so it is not the case that one outlier is causing the problem.

In light of these results, a second experiment, this time using budget shares, is undertaken. A budget share is defined as a specific expenditure (e.g., shelter and utilities) divided by total expenditures. Now, levels of incomes (not shares) are regressed on budget shares and other characteristics. Budget shares are used as independent variables for two reasons. First, income is the variable to be predicted; hence it cannot be on both sides of the equation in any way. Second, total expenditures can be used as a proxy for permanent income, as described earlier. In these equations the level of income remains transformed, but neither the shares nor their components (specific or total expenditures) are transformed in any way.

Once again, shares are not found to be useful in predicting income. For the self-employed singles, no expenditure share has a parameter estimate with a t-statistic indicating statistical significance at the 95% confidence level. Furthermore, regardless of income source, only the share of food at home has a larger R^2 value than its level-of-expenditure-regression counterpart, described later. Even in this case, the difference in R^2 is negligible – less than 0.01 regardless of income source. Perhaps budget shares do not predict as well as expenditures because the relationship between income and expenditures may not be linear. For example, the share of expenditures spent on shelter and utilities may be decreasing, but at a non-constant rate (i.e., $\partial^2 S/\partial I^2 \neq 0$, where S is the budget share and I is income). If this is the case, scatter plots of income (both transformed and actual) as a function of shares provide no clues as to what relationship might be plausible; however, linear relationships are evident in many plots of transformed income on transformed expenditures. Results of experiments in which transformed incomes are regressed directly on transformed expenditures and demographic characteristics are described in the text.

Appendix B: Labor variables

Although other authors cited in the text have used labor-related variables in their studies of Consumer Expenditure Survey income data, none has used variables

as detailed as those described in the text. Because hours per year worked and other similar variables are undoubtedly strong predictors of income, it is useful to examine their roles in the models.

For salaried singles only FULLTIME is not statistically significant (Table 8). The parameter estimate for HOURYEAR has a t-statistic of 22.0, higher than any other variable tested except total expenditures. The coefficients for OVERTIME and OTSLOPE also have large t-values, ranging from 8.0 to 10.0 depending on the model considered. The coefficients for HOURYEAR and OVERTIME are positive, while the coefficient for OTSLOPE is negative. This implies that those who work more than full-time receive some extra base pay, such as a bonus, but receive a lower return to overtime hours worked than to regular hours worked. (An F-test shows the hypothesis that the sum of the coefficients for HOURYEAR and OTSLOPE is zero can be rejected with 95% confidence in most cases and more than 90% confidence in all cases.) But caution must be used in interpreting this result because the workers included earn wage and salary income. A person who earns a high salary but whose overtime hours vary (e.g., a lawyer, doctor, or accountant) may indeed receive a lower "effective" wage (i.e., salary divided by hours per year) than someone working fewer hours in another occupation. Unfortunately, with CE data there is no way to distinguish between wage earners and salary earners to test these ideas.

For the self-employed these variables are also significant (Table 9), though generally with smaller t-values than for the salaried singles (Table 8). However, the parameter estimate for FULLTIME is negative, implying that those who work exactly 2,080 hours per year earn less, if all else is equal, than those who do not. For those who work more than full-time this result is not so surprising; they may be aggressively seeking business or once again may be in high paid professions that require many overtime hours. For those who work less than full-time, the result is puzzling - those who work less are predicted to have higher incomes (all else being equal). This is perhaps because working full-time (i.e., 40 hours per week, 52 weeks per year) is common for so many workers. Indeed, 21 % of the self-employed singles sampled work exactly 2,080 hours per year. There is a large potential for variation at this point, since some persons are starting new businesses, some are established, some are having slow periods, and so forth, but all are working 2,080 hours per year. It is possible that those with new or slow businesses are pulling down the mean income for the full-timers compared to par and overtime workers.

The test of the hypothesis that reporting a second source of labor income indicates a lower than average value for the primary source of income yields results (Tables 8 and 9) that are inconclusive due to a lack of statistical significance for both salaried and self-employed singles. However, even if a strong evidence of a relationship were found to exist, the model does not account for the type of relationship that may exist. For example, although the sign of the parameter estimate for the selfemployed is consistently negative, providing weak evidence that the self-employed

singles use wage and salary income to supplement lower earnings, or vice versa. it also may be true that the dual-source self-employed singles have left a salaried position to start a business, and therefore report income from wages and salaries and also lower self-employed earnings than their more established counterparts.

Appendix C: Variable description

Dependent variables

BOXSELF: Sum of self-employment income (business or farm), divided by CPI for month of interview, and subjected to Box-Cox transformation.

BOXWGSAL: Wage and salary income, divided by CPI for month of interview. and subjected to Box-Cox transformation.

Expenditure variables

Note: All expenditure variables are divided by CPI for month of interview, and subjected to Box-Cox transformation.

BOXEXP: Total expenditures:

BOXFOODH: Food at home.

BOXSHELU: Shelter (rent or owned dwelling expenditures for primary home) and utilities.

BOXTELE: Telephone services.

BOXBASIC: Basic goods and services (food at home, shelter and utilities, apparel and services).

BOXRLFUN: Recreation and related expenditures (entertainment, food away from home, lodging away from home).

Other independent variables

AGE: Age of the respondent.

AGESQ: Squared age of the reference person.

EDUCLEVL: Educational attainment of the respondent, with 0 being no schooling and 18 being at least 2 years of graduate school.

AGEEDUC: AGE * EDUCLEVL.

AGESQED: AGESQ * EDUCLEVL.

TM-INTER: Length of interview in minutes.

HOURYEAR: Number of hours per year worked.

FULLTIME: Dummy variable; equals one if HOURYEAR equals 2,080.

OVERTIME: Dummy variable; equals one if HOURYEAR exceeds 2,080.

OTSLOPE: OVERTIME * HOURYEAR.

OCCUPATIONAL CLASSES: TECHSALE: Respondent is in technical/sales work.

SERVICES: Respondent is in service work.
Control group is managers and professionals.

OTHLBINC: Dummy variable; indicates secondary source of labor income.

BLACK: Respondent is black.

FEMALE: Respondent is female.

BLACKFEM: BLACK * FEMALE.

STUDENT: Respondent is enrolled in college full- or part-time.

REGION OF RESIDENCE: NOREAST/MIDWEST/WEST: Indicate region in which consumer unit is located. Control group is located in Southern region.

RURAL: Consumer unit is located in a rural area.

RENTER: Respondent rents primary dwelling.

OWNOMORT: Respondent owns primary dwelling outright (i.e., no mortgage).

NOMBEXP: OWNOMORT * BOXEXP. Note: This variable is redefined and re-named as appropriate. For example, when food at home expenditures are included in the income prediction models, this variable is called "NOMRTFDH", and equals OWNOMORT times BOXFOODH.

SEASON OF INTERVIEW: QUARTER2/QUARTER3/QUARTER4: Indicate in which part of the year the interview takes place. Control group is the first quarter of the year (January, February, or March).

RECESS: Interview took place in 1991 or 1992. (Appears in self-employed regressions only.)

Table 8
Results of wage and salary income data regressed on expenditures and other characteristics: parameter estimates and t-statistics

	No expenditures		Total expenditures		Food at home		Shelter & utilities		Telephone services		Basic goods & services		Recreation & related expenditures	
Variable names	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value
R^2	0.6498	N/A	0.7091	N/A	0.6512	N/A	0.6690	N/A	0.6537	N/A	0.6763	N/A	0.6687	N/A
INTERCEPT	26.455	1.339	−24.942	−1.372	24.419	1.237	38.749	2.011	36.352	1.840	30.380	1.598	15.319	0.796
EXPENDITURES	N/A	N/A	5.344	20.804	0.158	2.848	0.302	11.231	0.411	4.896	0.886	13.331	1.242	11.067
AGE_REF	0.412	−0.440	−1.072	−1.250	0.338	0.361	−0.419	−0.458	−0.097	−0.104	−0.692	−0.765	0.474	0.520
AGESQ	−0.003	−0.268	0.010	1.158	−0.002	−0.209	0.003	0.360	0.002	0.237	0.006	0.654	−0.003	−0.367
EDUCLEVL	−1.006	−0.697	−3.117	−2.361	−0.977	−0.677	−2.219	−1.573	−1.788	−1.237	−2.550	−1.830	−1.180	−0.840
AGEEDUC	0.136	1.969	0.196	3.109	0.133	1.931	0.163	2.425	0.164	2.384	0.178	2.681	0.130	1.939
AGESQED	−0.001	−2.202	−0.002	−3.129	−0.001	−2.156	−0.001	−2.436	−0.001	−2.581	−0.001	−2.682	−0.001	−2.111
TM_INTER	0.090	5.606	0.031	2.082	0.089	5.538	0.072	4.590	0.080	4.951	0.064	4.130	0.066	4.171
HOURYEAR	0.026	22.897	0.022	21.639	0.026	22.912	0.024	21.926	0.025	22.491	0.024	22.245	0.025	22.870
FULLTIME	−0.564	−0.381	−0.204	−0.151	−0.480	−0.325	−0.267	−0.183	−0.354	−0.240	−0.303	−0.213	−0.660	−0.458
OVERTIME	48.746	9.054	42.141	8.566	48.154	8.952	45.541	8.683	48.027	8.963	45.430	8.762	48.277	9.215
OTSLOPE	−0.022	−10.106	−0.019	−9.938	−0.021	−10.028	−0.020	−9.806	−0.021	−9.989	−0.021	−10.020	−0.022	−10.398
TECHSALE	−6.531	−5.713	4.374	−4.167	−6.271	−5.476	−6.253	−5.615	−6.510	−5.724	−5.588	−5.060	−5.713	−5.121
PRECPROD	−8.397	−4.139	−5.270	−2.838	−8.562	−4.223	−7.534	−3.815	−8.206	−4.065	−7.358	−3.765	−6.313	−3.184
OPERATOR	−11.818	−7.815	−7.388	−5.291	−11.754	−7.779	−10.479	−7.089	−11.226	−7.433	−9.881	−6.742	−9.776	−6.586
SERVICES	−13.471	−9.377	−9.003	−6.730	−13.272	−9.195	−12.722	−9.028	−13.470	−9.383	−11.702	−8.353	−11.620	−8.186
OTHLBINC	−1.194	−0.546	−1.413	−0.708	−0.977	−0.447	−0.607	−0.285	−1.406	−0.646	−0.245	−0.116	−2.003	−0.940
BLACK	−2.665	−1.284	0.367	0.194	−2.768	−1.335	−2.189	−1.084	−2.265	−1.096	−2.659	−1.332	−0.872	−0.430
FEMALE	−6.365	−6.503	−3.737	−4.146	−6.173	−6.304	−6.087	−6.394	−6.620	−6.789	−6.345	−6.741	−4.799	−4.985
BLACKFEM	5.635	1.965	2.968	1.134	5.670	1.981	3.930	1.408	4.736	1.657	4.054	1.469	5.819	2.086
STUDENT	−6.390	−4.877	−4.300	−3.585	−6.123	−4.666	−4.638	−3.613	−6.425	−4.929	−4.158	−3.270	−6.308	−4.948
NOREAST	4.554	3.557	4.211	3.604	4.421	3.455	4.254	3.412	4.655	3.654	3.827	3.102	4.388	3.521

Table 8
(Continued)

Variable names	No expenditures		Total expenditures		Food at home		Shelter & utilities		Telephone services		Basic goods & services		Recreation & related expenditures	
	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value
MIDWEST	0.820	0.668	1.080	0.965	1.036	0.844	0.747	0.626	0.771	0.632	0.981	0.832	0.771	0.646
WEST	1.385	1.148	1.237	1.125	1.307	1.085	0.657	0.560	1.569	1.307	0.442	0.381	1.252	1.068
RURAL	−9.140	−5.822	−6.959	−4.841	−9.019	−5.740	−7.966	−5.200	−8.942	−5.722	−7.261	−4.776	−8.966	−5.866
RENTER	−11.590	−9.876	−6.463	−5.877	−11.482	−9.793	−8.942	−7.673	−10.887	−9.252	−8.691	−7.556	−9.697	−8.394
OWNOMORT	−12.282	−6.566	30.953	2.614	−12.936	−2.417	6.414	1.351	−11.995	−2.632	10.241	1.400	−4.363	−1.273
NOMRT*EXP	N/A	N/A	−2.117	−3.214	0.023	0.120	−0.327	−3.174	0.010	0.028	−0.575	−2.488	−0.701	−2.416
QUARTER2	1.409	1.193	1.505	1.399	1.512	1.283	1.491	1.299	1.304	1.110	2.077	1.827	1.408	1.225
QUARTER3	−1.956	−1.626	−1.934	−1.764	−1.962	−1.634	−1.399	−1.195	−2.052	−1.715	−0.985	−0.851	−1.883	−1.609
QUARTER4	−0.625	−0.510	−1.881	−1.679	−0.596	−0.487	−0.697	−0.584	−0.826	−0.677	−0.534	−0.453	−0.675	0.565

NOMRT * EXP: Interaction of owned home, no mortgage and expenditures.

Table 9
Results of self-employment income data regressed on expenditures and other characteristics: parameter estimates and t-statistics

	No expenditures		Total expenditures		Food at home		Shelter & utilities		Telephone services		Basic goods & services		Recreation & related expenditures	
Variable names	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value
R^2	0.6042	N/A	0.7070	N/A	0.6167	N/A	0.6422	N/A	0.6209	N/A	0.6617	N/A	0.6248	N/A
INTERCEPT	31.013	1.539	5.3040	0.299	25.170	1.245	34.695	1.797	38.407	1.919	24.094	1.283	24.604	1.239
EXPENDITURES	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
AGE_REF	−0.645	−0.894	−0.679	−1.077	−0.535	−0.736	−0.874	−1.257	−0.971	−1.347	−0.719	−1.066	−0.466	−0.654
AGESQ	0.003	0.556	0.004	0.909	0.002	0.416	0.005	0.892	0.006	1.016	0.004	0.736	0.002	0.401
EDUCLEVL	−0.834	−0.574	−0.949	−0.746	−0.582	−0.398	−1.385	−0.987	−1.529	−1.051	−1.057	−0.778	−0.317	−0.220
AGEEDUC	0.0457	0.855	0.042	0.909	0.035	0.654	0.058	1.143	0.068	1.281	0.046	0.928	0.025	0.478
AGESQED	−0.000	−0.634	−0.000	−0.818	−0.000	−0.427	−0.000	−0.887	−0.000	−1.055	−0.000	−0.689	−0.000	−0.308
TM_INTER	0.021	1.449	0.013	1.039	0.020	1.410	0.017	1.236	0.023	1.615	0.017	1.278	0.020	1.447
HOURYEAR	0.006	5.444	0.005	5.947	0.005	5.137	0.006	5.585	0.006	5.538	0.005	5.441	0.006	5.688
FULLTIME	−4.534	−2.682	−4.298	−2.928	−3.991	−2.342	−4.933	−3.046	−4.552	−2.733	−4.461	−2.838	−4.422	−2.666
OVERTIME	11.859	2.676	9.838	2.559	11.055	2.503	10.357	2.435	12.148	2.783	9.769	2.358	12.129	2.794
OTSLOPE	−0.006	−3.472	−0.005	−3.786	−0.005	−3.162	−0.005	−3.464	−0.006	−3.600	−0.005	−3.304	−0.006	−3.653
TECHSALE	−2.564	−2.105	−2.787	−2.643	−2.509	−2.076	−2.578	−2.208	−2.928	−2.384	−2.470	−2.172	−2.548	−2.135
PRECPROD	−2.634	−1.510	−2.007	−1.322	−2.693	−1.558	−2.467	−1.477	−2.288	−1.310	−2.488	−1.530	−2.686	−1.571
OPERATOR	−1.212	−0.865	0.175	0.143	−1.132	−0.815	−0.610	−0.452	−0.854	−0.607	−0.688	−0.526	−0.960	−0.698
SERVICES	−4.013	−2.541	−2.285	−1.650	−3.965	−2.535	−3.458	−2.281	−4.038	−2.590	−3.355	−2.274	−3.878	−2.488
OTHLBINC	−1.968	−1.356	−2.072	−1.649	−1.736	−1.203	−1.858	−1.330	−1.882	−1.317	−2.013	−1.489	−2.363	−1.654
BLACK	4.497	1.772	2.970	1.345	3.990	1.581	3.881	1.595	3.892	1.543	3.612	1.524	5.154	2.064
FEMALE	0.207	0.177	−0.224	−0.221	−0.001	−0.002	−0.440	−0.391	−0.252	−0.216	−0.682	−0.621	0.639	0.554
BLACKFEM	−6.628	−1.530	−5.087	−1.348	−7.366	−1.713	−5.124	−1.228	−6.060	−1.416	−6.307	−1.563	−8.046	−1.882
STUDENT	−2.491	−1.727	−1.381	−1.097	−1.814	−1.224	−1.017	−0.701	−2.041	−1.427	−0.718	−0.510	−2.533	−1.788
NOREAST	0.832	0.600	−0.920	−0.752	0.375	0.268	−0.457	−0.333	0.214	0.154	−0.683	−0.511	0.812	0.598

Table 9
(Continued)

Variable names	No expenditures		Total expenditures		Food at home		Shelter & utilities		Telephone services		Basic goods & services		Recreation & related expenditures	
	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value	Param. est.	t-value
MIDWEST	2.280	1.839	2.251	2.097	2.582	2.088	2.031	1.698	2.260	1.851	2.455	2.121	2.287	1.883
WEST	2.826	2.425	2.249	2.205	2.732	2.353	1.688	1.429	2.517	2.175	1.882	1.676	2.835	2.474
RURAL	−0.820	−0.595	−1.057	−0.882	−1.328	−0.951	−0.395	−0.299	−1.251	−0.914	−0.662	−0.515	−0.860	−0.634
RENTER	0.002	0.002	0.961	0.955	0.201	0.180	0.942	0.834	0.105	0.095	1.068	0.979	−0.037	−0.034
OWNOMORT	1.542	1.047	−11.901	−1.586	0.673	0.147	−1.173	−0.330	0.562	0.201	−4.346	−0.796	−2.912	−1.195
NOMRT*EXP	N/A	N/A	0.726	1.812	0.038	0.187	0.132	1.217	0.063	0.306	0.353	1.346	0.274	2.102
QUARTER2	−0.666	−0.514	−0.107	−0.094	−0.418	−0.316	−0.598	−0.481	0.264	0.200	−0.122	−0.100	−1.053	−0.814
QUARTER3	−1.606	−1.292	−0.738	−0.679	−1.510	−1.214	−1.826	−1.533	−1.465	−1.193	−1.685	−1.456	−1.977	−1.603
QUARTER4	−2.542	−1.955	−1.021	−0.887	−2.581	−1.950	−2.336	−1.875	−2.088	−1.617	−2.062	−1.691	−2.276	−1.764
RECESS	−0.133	−0.138	−0.506	−0.601	−0.318	−0.332	−0.571	−0.611	−0.627	−0.647	−0.501	−0.554	−0.333	−0.352

NOMRT * EXP: Interaction of owned home, no mortgage and expenditures.

Salaried singles: descriptive statistics for variables used in regressions

Variable	n	Mean (weighted)	Minimum value	Maximum value	STD error of mean[a]	Sum
Demographic characteristics						
AGE_REF	2207	35.9333	17.00000	83.00	0.300	3.6414411E+08
AGESQ	2207	1490.3087	289.00000	6889.00	25.493	1.5102641E+10
EDUCLEVL	2207	13.9802	0.00000	18.00	0.057	1.4167410E+08
AGEEDUC	2207	493.7117	0.00000	1440.00	4.235	5.0032256E+09
AGESQED	2207	20039.2677	0.00000	115200.00	334.623	2.0307595E+11
TM_INTER	2207	64.5619	3.00000	220.00	0.619	6.5426388E+08
HOURYEAR	2207	1826.1561	4.00000	4680.00	17.644	1.8506085E+10
Hours/Week	2207	40.0948	2.00000	90.00	0.267	4.0631660E+08
Weeks/Year	2207	44.3677	1.00000	52.00	0.291	4.4961805E+08
FULLTIME	2207	0.2812	0.00000	1.00	0.010	2.8491483E+06
OVERTIME	2207	0.2808	0.00000	1.00	0.010	2.8455035E+06
OTSLOPE	2207	761.3508	0.00000	4680.00	26.400	7.7154533E+09
TECHSALE	2207	0.3065	0.00000	1.00	0.010	3.1059446E+06
PRECPROD	2207	0.0600	0.00000	1.00	0.005	6.0798774E+05
OPERATOR	2207	0.1485	0.00000	1.00	0.008	1.5049511E+06
SERVICES	2207	0.1518	0.00000	1.00	0.008	1.5383977E+06
OTHLBINC	2207	0.0407	0.00000	1.00	0.004	4.1247785E+05
BLACK	2207	0.1005	0.00000	1.00	0.006	1.0180811E+06
FEMALE	2207	0.4616	0.00000	1.00	0.011	4.6775417E+06
BLACKFEM	2207	0.0519	0.00000	1.00	0.005	5.2594777E+05
STUDENT	2207	0.2231	0.00000	1.00	0.009	2.2606835E+06
NOREAST	2207	0.2147	0.00000	1.00	0.009	2.1753238E+06
MIDWEST	2207	0.2580	0.00000	1.00	0.009	2.6150363E+06
SOUTH[b]	2207	0.2815	0.00000	1.00	0.010	2.8528226E+06
WEST	2207	0.2458	0.00000	1.00	0.009	2.4907183E+06
RURAL	2207	0.0857	0.00000	1.00	0.006	8.6830393E+05
RENTER	2207	0.7030	0.00000	1.00	0.010	7.1240214E+06
OWNOMORT	2207	0.0864	0.00000	1.00	0.006	8.7541949E+05
QUARTER1[b]	2207	0.2718	0.00000	1.00	0.009	2.7540975E+06
QUARTER2	2207	0.2589	0.00000	1.00	0.009	2.6235812E+06
QUARTER3	2207	0.2433	0.00000	1.00	0.009	2.4653228E+06
QUARTER4	2207	0.2261	0.00000	1.00	0.009	2.2908995E+06
Expenditure variables (divided by CPI)						
Total exps.[c]	2207	14944.9054	654.14002	209578.95	237.622	1.5145019E+11
Food at home	2207	328.2091	0.00000	4193.88	5.603	3.3260387E+09
Shelter/util.	2207	979.5326	0.00000	6924.84	15.736	9.9264862E+09
Telephone	2207	95.0666	0.00000	1570.40	2.096	9.6339543E+08
Basics	2207	1526.5115	0.00000	12808.41	22.033	1.5469516E+10
Recreation	2207	521.7984	0.00000	15699.22	20.080	5.2878537E+09
Box–Cox transformations						
BOXEXP	2207	18.0095	9.99073	29.00	0.045	1.8250604E+08
BOXFOODH	2207	23.4407	−2.35294	79.16	0.182	2.3754590E+08
BOXSHELU	2207	49.2783	−2.10526	138.34	0.425	4.9938143E+08
BOXTELE	2207	10.2545	−2.66667	39.45	0.121	1.0391803E+08
BOXBASIC	2207	32.6774	−2.85714	75.41	0.172	3.3114949E+08
BOXRLFUN	2207	10.4707	−5.00000	29.53	0.094	1.0610899E+08

[a]The standard error of the mean is calculated to be $s/n^{0.5}$, where $s^2 = \sum w_i(x_i - X_w)^2 / \sum w_i$, where w_i is the population weight for the i-th observation for variable x; x_i is the value of the i-th observation of x; X_w is the weighted mean of x; and n is the sample size.

[b]Control group variable in regression; values shown here for convenience.

[c]Total quarterly expenditures multiplied by four to annualize for easier comparison to income data. All other expenditures are in quarterly form.

Variable	n	Mean (weighted)	Minimum value	Maximum value	STD error of mean[a]	Sum
Interaction of Box–Cox transformations with owned home, no mortgage						
Total exps.[c]	2207	1.5196	0.00000	25.79	0.106	1.5399336E+07
Food at home	2207	2.1958	−2.35294	63.93	0.160	2.2251549E+07
Shelter/util.	2207	3.5749	0.00000	90.54	0.263	3.6227420E+07
Telephone	2207	0.9081	−2.66667	20.27	0.067	9.2029008E+06
Basics	2207	2.6015	0.00000	52.93	0.184	2.6363828E+07
Recreation	2207	0.8334	−5.00000	23.29	0.066	8.4454786E+06
Income values (Means for those reporting. Real indicates original value is divided by CPI.)						
Wage/salary	2207	19546.6633	6.00000	375000.00	415.612	1.9808395E+11
Real wage/sal.	2207	15727.2585	4.68750	290247.68	331.410	1.5937848E+11
BOXWGSAL	2207	87.7270	2.09295	295.53	0.715	8.8901627E+08
Business	67	2727.9686	−35000.00000	55000.00	1288.260	8.8067340E+08
Farm	8	2739.2412	−5000.00000	8200.00	1285.686	1.0637416E+08
Other variables						
CPI	2207	124.1431	115.70000	133.80	0.114	1.2580535E+09
Pop. weight[d]	2207	4591.7086	199.01517	18295.40	49.772	1.0133901E+07

[a]The standard error of the mean is calculated to be $s/n^{0.5}$, where $s^2 = \sum w_i(x_i - X_w)^2 / \sum w_i$, where w_i is the population weight for the i-th observation for variable x; x_i is the value of the i-th observation of x; X_w is the weighted mean of x; and n is the sample size.

[c]Total quarterly expenditures multiplied by four to annualize for easier comparison to income data. All other expenditures are in quarterly form.

[d]Values for this variable (population weight for each observation) are unweighted.

Self-employed singles: descriptive statistics for variables used in regressions

Variable	n	Mean (weighted)	Minimum value	Maximum value	STD error of mean[a]	Sum
Demographic characteristics						
AGE_REF	202	47.7447	18.00000	92.00	1.27	4.7042261E+07
AGESQ	202	2607.3266	324.00000	8464.00	132.17	2.5689678E+09
EDUCLEVL	202	13.7187	2.00000	18.00	0.23	1.3516845E+07
AGEEDUC	202	637.7585	82.00000	1368.00	18.12	6.2837581E+08
AGESQED	202	33738.2580	3362.00000	118496.00	1633.30	3.3241903E+10
TM_INTER	202	68.7011	15.00000	210.00	2.23	6.7690407E+07
HOURYEAR	202	1683.9985	20.00000	4680.00	74.38	1.6592236E+09
Hours/Week	202	38.8445	2.00000	90.00	1.24	3.8273023E+07
Weeks/Year	202	42.0821	1.00000	52.00	1.11	4.1462958E+07
FULLTIME	202	0.1896	0.00000	1.00	0.03	1.8682684E+05
OVERTIME	202	0.2571	0.00000	1.00	0.03	2.5333336E+05
OTSLOPE	202	779.1944	0.00000	4680.00	95.83	7.6773091E+08
TECHSALE	202	0.2185	0.00000	1.00	0.03	2.1525907E+05
PRECPROD	202	0.0901	0.00000	1.00	0.02	8.8789822E+04
OPERATOR	202	0.2163	0.00000	1.00	0.03	2.1307002E+05
SERVICES	202	0.1312	0.00000	1.00	0.02	1.2924433E+05
OTHLBINC	202	0.1180	0.00000	1.00	0.02	1.1627950E+05
BLACK	202	0.0480	0.00000	1.00	0.02	4.7261492E+04
FEMALE	202	0.3168	0.00000	1.00	0.03	3.1209166E+05
BLACKFEM	202	0.0166	0.00000	1.00	0.01	1.6385800E+04
STUDENT	202	0.1353	0.00000	1.00	0.02	1.3333283E+05
NOREAST	202	0.1735	0.00000	1.00	0.03	1.7090327E+05
MIDWEST	202	0.2213	0.00000	1.00	0.03	2.1806527E+05
SOUTH[b]	202	0.3203	0.00000	1.00	0.03	3.1556148E+05
WEST	202	0.2850	0.00000	1.00	0.03	2.8075805E+05
RURAL	202	0.1410	0.00000	1.00	0.02	1.3891401E+05
RENTER	202	0.4986	0.00000	1.00	0.04	4.9130510E+05
OWNOMORT	202	0.2313	0.00000	1.00	0.03	2.2793560E+05
QUARTER1[b]	202	0.2603	0.00000	1.00	0.03	2.5642565E+05
QUARTER2	202	0.2591	0.00000	1.00	0.03	2.5525123E+05
QUARTER3	202	0.2427	0.00000	1.00	0.03	2.3908599E+05
QUARTER4	202	0.2380	0.00000	1.00	0.03	2.3452520E+05
RECESS	202	0.3937	0.00000	1.00	0.03	3.8789003E+05
Expenditure variables (divided by CPI)						
Total exps.[c]	202	16700.1001	2392.55237	68950.46	776.64	1.6454409E+10
Food at home	202	365.4856	0.00000	2022.47	21.29	3.6010863E+08
Shelter/util.	202	1096.2962	1.56250	5380.59	62.32	1.0801676E+09
Telephone	202	115.7552	0.00000	717.92	7.69	1.1405224E+08
Basics	202	1661.3813	120.46444	5953.74	76.92	1.6369392E+09
Recreation	202	648.0917	0.00000	7700.25	62.76	6.3855698E+08
Box–Cox transformations						
BOXEXP	202	18.3802	13.15669	24.20	0.15	1.8109824E+07
BOXFOODH	202	20.0170	−2.66667	43.64	0.49	1.9722495E+07
BOXSHELU	202	31.4277	0.48581	64.17	0.77	3.0965349E+07
BOXTELE	202	11.4699	−2.66667	28.74	0.40	1.1301189E+07
BOXBASIC	202	20.5279	9.25180	31.14	0.29	2.0225931E+07
BOXRLFUN	202	14.5529	−3.63636	38.97	0.55	1.4338765E+07

[a]The standard error of the mean is calculated to be $s/n^{0.5}$, where $s^2 = \sum w_i(x_i - X_w)^2 / \sum w_i$, where w_i is the population weight for the i-th observation for variable x; x_i is the value of the i-th observation of x; X_w is the weighted mean of x; and n is the sample size.

[b]Control group variable in regression; values shown here for convenience.

[c]Total quarterly expenditures multiplied by four to annualize for easier comparison to income data. All other expenditures are in quarterly form.

Variable	n	Mean (weighted)	Minimum value	Maximum value	STD error of mean[a]	Sum
Interaction of Box–Cox transformations with owned home, no mortgage						
Total exps.[c]	202	4.1795	0.00000	24.20	0.54	4.1180166E+06
Food at home	202	4.7305	0.00000	39.40	0.63	4.6609260E+06
Shelter/util.	202	6.4439	0.00000	50.44	0.88	6.3490916E+06
Telephone	202	2.7203	0.00000	28.74	0.34	2.6802686E+06
Basics	202	4.4420	0.00000	30.78	0.58	4.3766293E+06
Recreation	202	3.2757	−3.63636	30.92	0.50	3.2275290E+06
Income values (Means for those reporting. Real indicates original value is divided by CPI.)						
Business	188	17865.1069	100.00000	268618.00	1790.00	1.6044000E+10
Farm	15	17238.0755	100.00000	100000.00	6960.53	1.5740765E+09
Real self-emp.	202	13872.4929	74.18398	193808.08	1324.44	1.3668402E+10
BOXSELF	202	24.5975	6.83129	52.07	0.60	2.4235620E+07
Wage/salary	23	3287.3582	250.00000	14000.00	651.06	3.6597529E+08
Other variables						
CPI	202	129.8007	115.70000	142.00	0.59	1.2789107E+08
Pop. weight[d]	202	4877.6637	221.87894	17634.81	187.76	9.8528807E+05

[a]The standard error of the mean is calculated to be $s/n^{0.5}$, where $s^2 = \sum w_i(x_i - X_w)^2 / \sum w_i$, where w_i is the population weight for the i-th observation for variable x; x_i is the value of the i-th observation of x; X_w is the weighted mean of x; and n is the sample size.
[c]Total quarterly expenditures multiplied by four to annualize for easier comparison to income data. All other expenditures are in quarterly form.
[d]Values for this variable (population weight for each observation) are unweighted.

Acknowledgements

The authors gratefully acknowledge Donald B. Rubin (Chairman, Department of Statistics, Harvard University) and David Brownstone (Department of Economics. University of California) for discussing an earlier version of this paper at the Joint Statistical Meetings of the American Statistical Association (Toronto, Ontario, Canada, August 1994); Charles H. Alexander, Jr. (U.S. Bureau of the Census) for his help and support throughout the writing of this paper; and Ralph Bradley (U.S. Bureau of Labor Statistics) for his insightful comments, especially on the final drafts of the paper.

References

[1] G. Bannock, R. Baxter and R. Rees, A *Dictionary of Economics*, Penguin Books Ltd., Harmondsworth, Middlesex, England, 1972.

[2] JR. Blaylock and W.N. Blisard, Wine consumption by U.S. men, *Applied Economics, 24* May, 1993, 64.5451.

[3] G.E.P. Box and D.R. Cox, An analysis of transformations, J. R. *Stat Soc.*, Series B. 1964. *211-243.*

[4] D. Brownstone and P. Englund, The demand for housing in Sweden: equilibrium choice of tenure and type of dwelling, J. *Urban Econ.* 29(3) (1991), 267-281.

[5] N. Chand and C.H. Alexander, Unpublished results of recent preliminary experiments using data with a log transformation. For a description of their methodology, see: Chand and Alexander, Imputing income for an N-person consumer unit, 1994 *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Alexandria, Virginia, 199-1, pp. 412-.116..

(6] M. David, R.J.A. Little, ME. Samuhel and R.K. Triest, Alternative methods for CPS income imputation, J. *Am. Star. Assoc.* (Applications) 81(393) (1986), 29-41.

[7] J.L. Eltinge and 1.5. Yansaneh, Weighting adjustments for income nonresponse in the U.S. Consumer Expenditure Survey. Technical Report No. 202, Texas A&M University. Department of Statistics, 1993.

[8] *R.* Gillingham and *R.* Hagemann, Cross-sectional estimation of a simultaneous model of tenure choice and housing services demand, J. *Urban Econ.* 14(1) (1983), 16-39.

[9] 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, J. *Am. Star. Assoc.* 77(378) (1982), 251-261.

[10] P. Kennedy. *A Guide to Econometrics,* 3rd ed., MIT Press, Cambridge, 1992.

[11] E.L. Lehmann, *Nonparametrics: Statistical Methods Based nn Ranks,* Holden-Day, San Francisco, 1975.

[12] L. Lillard, J.P. Smith and F. Welch, What do we really know about wages? The importance of nonreporting and Census imputation, J. *Polit. Econ.* 94(3) (1986), 489-506.

[13] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Dana,* John Wiley & Sons, New York, 1987.

[14] G.D. Paulin, A comparison of consumer expenditures by housing tenure, J. *Cons. Affairs 29(1)* (Summer, 1995), 164-198.

[IS] G.D. Paulin and D.L. Ferraro, Imputing income in the Consumer Expenditure Survey, *Mon. Labor Rev.,* December, 1994, 23-31.

[16] G.D. Paulin and E.M. Sweet, Modeling income in the U.S. Consumer Expenditure Survey, J. O/]: *Star.* (forthcoming).

(17] S. Scott and D.J. Rope, Distributions and transformations for family expenditures, 1993 *Proceedings of the Section on Social Statistics,* American Statistical Association, 1993, 741-746.

[18] *Survey of Current Business,* United States Department of Commerce, Economics and Statistics Administration, Bureau of Economic Analysis, 71(12) (December, 1991), Table 1.2, p. 3.