The Missing Income Problem in Analyses of Engel Functions

Oral Capps, Jr. and Hsiang-Tai Cheng

The empirical evidence from the extant literature in demand analysis points to the importance of income in food expenditure relationships. However, roughly 30 percent of all households in the 1977–78 Nationwide Food Consumption Survey do not report income figures. The focus of this paper is on the missing income problem in analyses of Engel functions. This analysis statistically links particular demographic attributes in affecting the probability of reporting income information. Additionally, several techniques to overcome the missing income problem, namely, regression imputation, the Heckman procedure, and item deletion, are discussed. Empirical evidence suggests that the Heckman procedure is statistically superior to item deletion, and that regression imputation and the Heckman procedure yield similar results.

Key words: Engel functions, food expenditure, missing data.

With appropriate recognition of household size and composition, location, and other characteristics, empirical evidence from the extant literature in demand analysis points to the importance of income in food expenditure relationships (Prais and Houthakker; Brown and Deaton; Ferber).[1] The relationship between income and food consumption can be expressed a number of ways; for example, the percentage of total income spent for food (average budget share), the fraction of each extra dollar of income spent to purchase food (marginal budget share), and, similarly, the percentage change in the quantity of food resulting from a unit percent change in income (income elasticity). Information from the marginal budget shares out of different types of income have been used to assess the relative impacts on consumption of different federal food program alternatives, such as the effect of cash income supplementation versus food stamp supplementation or the impacts on consumption of participation in the School Lunch Program; Head Start Program; Women, Infant, and Children Program; and Nutrition Program for the Elderly. Finally, income elasticities have been used to classify goods as inferior, normal, or luxury (superior), referring to relative consumption changes with changes in income.

A substantial portion of research studies that focus attention on food expenditure-income relationships (Engel or expenditure functions) either employ household budget data or cross-section data. Examples of such data sets include the 1977–78 U.S. Department of Agriculture (USDA) Nationwide Household Food Consumption Survey (NFCS) and the current Bureau of Labor Statistics (BLS) continuous Quarterly Consumer Expenditure Surveys. The income data in the respective surveys in particular provide a valuable source of information for the analysis of household behavior. However, despite the wealth of demographic and economic information from the respective data bases, the nonreporting of household income is a common and pervasive problem. In particular, in the 1977–78 NFCS approxi-

Oral Capps, Jr., and Hsiang-Tai Cheng are, respectively, an associate professor in the Departments of Agricultural Economics and Statistics, and a research assistant, Department of Agricultural Economics, Virginia Polytechnic Institute and State University.

This research work was sponsored by the S-165 Regional Project on U.S. Food Demand and Consumption Behavior under Virginia Tech Project No. 2124270.

The authors wish to gratefully acknowledge the comments made by three anonymous reviewers. Any remaining errors are the sole responsibility of the authors.

[1] In addition, agricultural economists have centered attention on developing sets of own-price and cross-price elasticities. However, with the emphasis on household budget data or cross-section data, a substantial portion of research studies, with respect to demand analysis, deals primarily with food expenditure-income relationships.

Western Journal of Agricultural Economics, 11(1): 31–39

mately 30% of all housekeeping households do not report income figures but do report socioeconomic information.[2] This rather large percentage of households not reporting income is in general representative of household budget or cross-section data bases available to agricultural economists.

In this light, the focus of this paper centers on the missing income problem in analyses of Engel functions for beverages, fats and oils, fruits, grains, meat and meat alternates, milk products, sugars and sweets, vegetables, miscellaneous foods, and the aggregate food at home. Importantly, this paper does not address the reliability of income data from household budget or cross-section surveys (see Atkinson and Micklewright). A full description of the various food groups is available upon request from the authors. The source of data for this investigation is the 1977–78 NFCS. Specifically, the objectives of this paper are threefold: (*a*) to determine the relationship among several socioeconomic factors and the nonreporting of household income, (*b*) to discuss several procedures to try to overcome the missing income problem in household budget or cross-section surveys, and (*c*) in empirical fashion to compare and contrast the various procedures. With the relationship in hand from objective *a,* it is possible to determine the probability (or likelihood) that a household with a given set of attributes will not report income. Consequently, the results of this paper can be used by officials from either BLS or USDA to improve the collection of vital household income information in future surveys. Objectives *b* and *c* are worthwhile because "there is no best solution for dealing with the problem of missing observations . . ." (Pindyck and Rubinfeld, p. 245).

The Nonreporting of Household Income: Link to Demographic Attributes

The following socioeconomic characteristics are hypothesized to influence the nonreporting of household income: (*a*) race of the household head, (*b*) region, (*c*) urbanization or population density, (*d*) occupation of the household head, (*e*) season, (*f*) education of the household manager, (*g*) employment status of the household manager, (*h*) age of the household manager, and (*i*) sex of the household manager. Each of the socioeconomic factors is a binary variable. Binary variables assume the value of zero or one depending upon the attainment or nonattainment of particular attributes.

The household manager is defined as the female head in households with both male and female heads present and in households with only the female head present. In all other cases, the household manager is the male head. The household manager is solely responsible for the completion of the survey form. The completion of the form, generally accomplished in a single time period, is under the supervision of the interviewer.

Prior information is insufficient for hypotheses concerning the impacts of the socioeconomic variables on the nonreporting of household income. The statistical model for this study is given by

$$\begin{aligned} INCR = {} & a_0 + a_1 RACHH + a_2 R1 + a_3 R2 \\ & + a_4 R3 + a_5 URB1 + a_6 URB2 \\ & + a_7 OCUHH + a_8 S1 + a_9 S2 \\ & + a_{10} S3 + a_{11} EDHM + a_{12} EMPHM \\ & + a_{13} AGHM + a_{14} SXHM + e \end{aligned}$$

where

INCR = 1 if the household fails to report income,
0 otherwise;

RACHH = 1 if the household head is nonwhite,
0 otherwise;

*R*1 = 1 if the household is located in Northeast,
0 otherwise;

*R*2 = 1 if the household is located in North Central,
0 otherwise;

*R*3 = 1 if the household is located in South,
0 otherwise;

*URB*1 = 1 if the household resides in a central city,
0 otherwise;

*URB*2 = 1 if the household resides in a suburban area,
0 otherwise;

OCUHH = 1 if the household head is a white collar worker (professional and technical, managers, officers, and proprietors),
0 otherwise;

[2] Housekeeping households are defined as those households with at least one person having eaten ten or more meals from the household supply during the survey period.

*S*1 = 1 if spring (April, May, June 1977),
0 otherwise;
*S*2 = 1 if summer (July, August, September 1977),
0 otherwise;
*S*3 = 1 if fall (October, November, December 1977),
0 otherwise;
EDHM = 1 if the household manager is at least high school graduate,
0 otherwise;
EMPHM = 1 if the household manager is employed either full time or part time,
0 otherwise;
AGHM = 1 if the household manager is less than 35 years of age,
0 otherwise;
SXHM = 1 if the household manager is female,
0 otherwise;

Table 1. Descriptive Statistics of the Variables in the Qualitative Choice Model

Variable	Mean	Standard Deviation
INCR	.2956	.4563
RACHH	.1444	.3515
*R*1	.2396	.4268
*R*2	.2537	.4351
*R*3	.3452	.4754
*URB*1	.2906	.4540
*URB*2	.3510	.4773
OCUHH	.6305	.4826
*S*1	.2197	.4140
*S*2	.2324	.4224
*S*3	.2726	.4453
EDHM	.7038	.4565
EMPHM	.3945	.4887
AGHM	.3616	.4804
SXHM	.9328	.2503

and *e* is the disturbance term of the statistical model. The reference household is defined as one wherein the household head is white and a blue collar worker, the household is located in the West and in a nonmetropolitan area, the season is winter (January, February, March 1978), and the household manager is not a high school graduate, unemployed, at least 35 years of age, and male.

The empirical analysis includes data from 13,787 housekeeping households. Because relevant data on socioeconomic factors were missing, 245 housekeeping households were eliminated from the analysis. Descriptive statistics of the variables are exhibited in table 1. The means of the binary variables reflect the proportions of households that fall into particular categories. For example, roughly 30% of the households do not report income information, 14% are nonwhite, and 34% are located in the South.

In the context of this study, a household either reports income information or does not. This decision is a linear function of socioeconomic characteristics. In view of the dichotomous nature of the dependent variable, this formulation constitutes a qualitative choice model. Two widely accepted qualitative dependent model formulations are the probit and the logit specifications (Amemiya). These formulations circumvent the problems of ordinary least squares estimation of the standard linear probability model.

The linear probability model suffers from three deficiencies. First, the variance of the disturbance term of the model is heteroscedastic, which results in loss of efficiency of the parameter estimates (Goldberger). Second, the distribution of the disturbance term is not normal (Pindyck and Rubinfeld). As such, the classical tests of significance are not applicable because the tests depend on the normality of the disturbance term. Third, and perhaps most important, the linear probability formulation allows predictions to fall outside the interval between 0 and 1, inconsistent with the interpretation of such predictions as probabilities (Amemiya; Pindyck and Rubinfeld).

The probit and logit specifications circumvent the difficulties of the linear probability model via the use of monotonic transformations of the original model to guarantee that predictions lie in the unit interval. For the probit formulation, the probability that the *i*th household fails to report income can be obtained from the standard normal cumulative distribution function. For the logit formulation, this probability can be obtained from the logistic cumulative distribution function.[3] Although the cumulative logistic probability function is basically similar in form to the cu-

[3] Particularly readable presentations of the two specifications may be found in Pindyck and Rubinfeld as well as in Amemiya. For the probit formulation,

$$P_i = \int_{-\infty}^{z_i} (2\pi)^{-1/2} \exp(-s^2/2)\, ds, \quad -\infty < z_i < \infty$$

where z_i is the linear function of the explanatory regressors in the model. For the logit formation,

$$P_i = e^{z_i}/(1 + e^{z_i}), \quad -\infty < z_i < \infty.$$

P_i represents the probability that the *i*th housekeeping household fails to report income.

Table 2. Maximum Likelihood Estimates for the Probit and Logit Analyses

	Probit Analysis		Logit Analysis	
Variable	Parameter Estimate (Standard Error)	Change in Probability[a]	Parameter Estimate (Standard Error)	Change in Probability[b]
RACHH	−.0537*[c] (.0349)	−.0184[d]	−.0859* (.0589)	−.0249[e]
*R*1	+.0102 (.0369)	+.0035	+.0182 (.0619)	+.0052
*R*2	+.1421* (.0364)	+.0488	+.2374* (.0607)	+.0690
*R*3	+.0563* (.0353)	+.0193	+.0946* (.0592)	+.0275
*URB*1	−.1303* (.0300)	−.0447	−.2170* (.0503)	−.0631
*URB*2	−.0294 (.0272)	−.0101	−.0460 (.0451)	−.0133
OCUHH	+.0582* (.0248)	+.0200	+.0974* (.0414)	+.0283
*S*1	−.2702* (.0328)	−.0928	−.4506* (.0551)	−.1311
*S*2	−.1123* (.0317)	−.0386	−.1854* (.0524)	−.0542
*S*3	−.0568* (.032)	−.0195	−.0952* (.0498)	−.0277
EDHM	−.1739* (.0261)	−.0597	−.2923* (.0432)	−.0850
EMPHM	−.0172 (.2385)	−.0059	−.0307 (.0397)	−.0089
AGHM	−.2663* (.0267)	−.0915	−.4461* (.0451)	−.1298
SEX	−.0137 (.0506)	−.0047	−.0253 (.0862)	−.0073
Intercept	−.2469* (.0718)		−.3864* (.1208)	

[a] At the sample means $z_i = -.5475$ $f(z_i) = .3434$ (value of standard normal probability density function).

[b] At the sample means $z_i = -.8903$ $f(z_i) = .2910$ (value of logistic probability density function).

[c] Asterisk indicates statistically significant at $\alpha = .05$ level.

[de] Parameter estimate times the value of the respective probability density function.

mulative normal function (Amemiya), both formulations are employed in this analysis.

The estimation of the model rests on the maximum likelihood technique (Amemiya; Pindyck and Rubinfeld). The maximum likelihood coefficients are consistent and asymptotically normally distributed. Consequently, conventional tests of significance are applicable, and likelihood ratio tests are the natural method of inference.

The maximum likelihood estimates of the probit and logit analyses are exhibited in table 2. The partial derivatives of the nonlinear probability functions evaluated at sample means appear under the column heading, change in probability. The signs and magnitudes of the respective estimates are very similar for the logit and probit specifications. For the reference household, the probability, according to the logit analysis, that a housekeeping household fails to report income is .4045; and this probability according to the probit analysis is .4025. Blue collar workers and nonwhite households have significantly lower probabilities of not reporting income than white collar workers and white households.

Households located in the North Central region and South have significantly higher probabilities of not reporting income than households located in the West. No statistically significant differences in failing to report income are apparent between households located in the Northeast and the West. Households located in central city areas have significantly lower probabilities of not reporting income than households located in nonmetropolitan areas. The probability of not reporting income is statistically the same for households located in suburban and nonmetropolitan areas.

The probability of failing to report income is lower in the spring, summer, and fall than in the winter. As defined in this study, the winter season corresponds to the time of year for filing income taxes. Interestingly, according to the empirical evidence, the probability of not reporting income information in the NFCS is highest during the traditional period of reporting income tax information in relation to other periods throughout the year.

Households wherein managers have at least high school education and are at most thirty-five years of age have lower probabilities of not reporting income than managers at least thirty-five years of age without high school education. Sex and employment status of the household manager have no discernible impacts on the probability of not reporting income information.

A measure of goodness-of-fit of the model involves the correct classification of households as either reporting income or not reporting income on the basis of the explanatory variable information. On the basis of the 50–50 classification scheme (Amemiya, p. 1503), slightly more than 70% of the households are correctly classified as reporting or not reporting income using the logit and probit specifications.

Procedures to Handle the Missing Income Problem

The analysis up to this point statistically links particular demographic attributes, notably race of the household head, geographic location, population density, occupation of the household head, as well as education and age of the household manager to affecting the probability of reporting income information. Although useful, the work thus far does not address the issue of how to handle missing income values in demand analyses.

Even the best designed surveys suffer from nonresponse, especially for quantitative items like household income. When observations appear to be missing at random, then eliminating the observations is a reasonable procedure. Under the assumption that the observations dropped are random, the ordinary least squares estimator is unbiased and consistent although not necessarily efficient. In practice, however, the assumption that missing observations are random is unlikely to be realistic. In fact, the previous analysis provides empirical evidence to indicate that the pattern of missing household income information is systematic. Consequently, deletion of households failing to report income information from empirical analyses may lead to statistical problems due to sample selection bias. In order to reduce this bias and consequently to alleviate the missing income problem, this paper focuses on two techniques in particular: (*a*) regression imputation and (*b*) the Heckman procedure.

A number of imputation procedures can be used to replace missing values resulting from nonresponse and/or invalid data. The three basic imputation procedures are (*a*) the cold deck imputation procedure, (*b*) the hot deck imputation procedure, and (*c*) the regression imputation procedure (Chapman; Cox). In cold deck imputation, household income values from some previous census or survey are substituted for missing data. Hunter and West employed this procedure by matching socioeconomic variables common to the NCFS and the Survey of Income and Education to obtain estimates of missing household income figures. The major drawbacks of this approach are data compatibility as well as programming and computational requirements. In hot deck imputation, survey records are first sorted by household identification number. When a missing income value is encountered, the last reported nonmissing income value is imputed for the missing response. Because of the ease of use and flexibility of implementation of the hot deck technique, it is the most commonly used item nonresponse imputation procedure. However, no probability mechanism is attached to the assignment of missing values and, moreover, the same responses may be used repeatedly to supply missing income information. Consequently, the hot deck imputation procedure is most appropriate when dealing with qualitative variables and when the missing data points are few in number.

Regression imputation is most appropriate when there exist variables which can be used to predict the missing income response. This imputation technique rests on the assumption that a linear relationship exists between household income and a set of regressor variables. This relationship is observed for the respondent data only, but the researcher believes that this relationship also exists for the data from nonrespondents. Missing income responses for the nonrespondent households can then be replaced by the predicted response. According to Cox, the use of this technique has a greater potential of producing imputed values which are closer to the true unknown value than other imputation approaches. However, the researcher should be aware that only mean values are imputed when using the regression technique. If household income is predicted based upon geographic location, race, and household size, for instance, then every household with missing income data but sharing these sociodemographic characteristics receives the same imputed value.

Regression imputation corresponds to the use of instrumental variables. Although this technique yields consistent estimates, error variances associated with missing observations are larger than the error variances associated with nonmissing observations. In this case, the use of weighted least squares is reasonable to adjust for the efficiency loss associated with the problem of heteroscedasticity (Glaser; Dagenais). Further, the use of regression imputation typically overstates the statistical significance of empirical results. The reason stems from the fact that preliminary regressions from the original data set are necessary to obtain imp.itations for the missing data (Glaser; Dagenais).

In order for nonresponse imputation to be worthwhile, the compensation for the reduction in bias must exceed the additional variability induced by imputation. Otherwise, in terms of the mean-squared error criterion, imputation reduces rather than increases the accuracy of survey estimates. The extant literature provides little guidance on when to impute and when not to impute (Chapman; Cox).

The Heckman procedure avoids imputation altogether to handle the missing value problem. According to Heckman, the sample selection bias that arises from using least squares is characterized as a specification error or omitted variable problem. When a subsample of data containing only nonmissing observations on household income is used for model estimation, the conditional expectation of the disturbance terms, in the general case, is nonzero. Therefore, parameter estimates derived from the selected sample omit the conditional expectation of the disturbance term as a regressor. Heckman subsequently proposes an estimator that amounts to estimating the nonzero conditional mean and using least squares including this variable as a regressor.

This procedure entails two stages. The first stage involves the use of probit analysis to determine the inverse of Mill's ratio for each household (λ_h) (Heckman, p. 479). The probit analysis employs all available observations; the dependent variable takes on the value of one if the household reports income and zero otherwise. The second stage involves the use of the estimated λ_h as a regressor in the original model specification. The appropriate estimation technique is either ordinary or generalized least squares, but the estimation involves only nonmissing observations. The OLS procedure produces consistent estimates, but the GLS procedure, when implementation is possible (see Heckman), improves the precision of the estimates.

The Heckman procedure allows the researcher to statistically test for sample selection bias. If the estimated coefficient associated with λ_h is significantly different from zero, then sample selection bias exists; in general, one cannot sign the direction of bias. Therefore, the common procedure of deleting missing income observations from the analysis is inappropriate statistically speaking. On the other hand, if the estimated coefficient associated with λ_h is not significantly different from zero, then deleting missing income observations from the analysis does not introduce bias into the remaining coefficient estimates. However, there is an efficiency loss caused by the use of the truncated sample. Because of the ability of the Heckman procedure to test for sample selection bias and because the procedure avoids imputation, this technique appears to be the preferred procedure to handle the missing value problem.

Empirical Investigation of Alternative Procedures

To compare and contrast the various procedures in empirical fashion, this section deals with the missing income problem in analyses of Engel functions for several food groups. The fundamental importance in this application is to determine the sensitivity of income coefficients in Engel functions from the use of several procedures to handle the missing income problem. The mathematical model form of the Engel specification is as follows:

$$EXP_{ih} = f(\text{LOG } INCB4TAX_h, R1, R2, R3, URB1, URB2, S1, S2, S3, RACHH, MEALS, \text{LOG } HSIZE_h),$$

where EXP_{ih} refers to weekly expenditure on the ith food group for the hth household, LOG $INCB4TAX_h$ refers to the logarithm of annual household income in dollars for the hth household, $R1$, $R2$, $R3$ are dummy variables previously defined with reference to geographic region, $URB1$ and $URB2$ are dummy variables previously defined with reference to urbanization, $S1$, $S2$, $S3$ are dummy variables previously defined with reference to seasonality, $RACHH$ is a zero–one variable previously defined with respect to race, $MEALS$ refers to the number of meals eaten from the household food supply per week, and LOG $HSIZE_h$ refers to the logarithm of household size for the hth household.

The demographic variables region, urbanization, seasonality, and race are in general common to Engel functions to account primarily for tastes and preferences. According to empirical evidence (e.g., Aitchinson and Brown), increases in food expenditures are rapid as income rises, but saturation levels are approached at relatively low levels of income. The logarithm of household income accounts for this possible nonlinear form of the Engel function. The logarithm of household size accounts for potential economies of size in food expenditure relationships (Price; Buse and Salathe). Finally, the

Table 3. Descriptive Statistics of the Dependent and Regressor Variables

	Truncated Sample (9,817 Observations)			Original Sample (14,000 Observations)		
	Percentage of Zero Observations	Mean	Standard Deviation	Percentage of Zero Observations	Mean	Standard Deviation
	(%)	($/week)		(%)	($/week)	
Dependent Variables						
Beverages	16.85	3.20	4.76	16.79	3.29	5.04
Fats and oils	3.40	1.41	1.17	3.55	1.44	1.19
Fruits	5.03	3.45	3.10	4.80	3.58	3.24
Grains	.15	5.66	4.42	.12	5.71	4.44
Meat and meat alternates	.05	17.61	12.52	.03	18.05	12.76
Milk products	.56	5.74	4.54	.60	5.86	4.68
Miscellaneous foods	9.23	2.07	2.41	8.71	2.12	2.45
Sugars and sweets	8.54	1.19	1.67	8.17	1.22	1.63
Vegetables	.75	5.49	3.86	.67	5.62	3.93
Food at home	.00	45.86	27.04	.00	46.93	27.73
Regressor Variables						
*R*1		.2518	.4349		.2464	.4308
*R*2		.2549	.4366		.2649	.4411
*R*3		.1848	.3889		.1803	.3844
*URB*1		.3208	.4676		.3079	.4615
*URB*2		.3214	.4679		.3310	.4704
*S*1		.2661	.4428		.2472	.4313
*S*2		.2444	.4305		.2499	.4329
*S*3		.2397	.4277		.2523	.4342
RACHH		.1545	.3622		.1485	.3555
LOG *INCB4TAX*		9.2161	.7517		9.2804[a]	.6538[a]
MEALS		57.4361	32.6664		58.0148	32.7841
LOG *HSIZE*		.9438	.5770		.9623	.5653
INCB4TAX		12,697.9	7,701.3		12,767.5[a]	6,687.41[a]

[a] With regression imputation.

inclusion of *MEALS* accounts for the number of meals eaten at home in the Engel function.

The analysis for this investigation includes data from usable schedules for 14,000 housekeeping households. Only 32 housekeeping households are excluded from the original set of 14,032; the reason for exclusion is the nonreporting of relevant demographic information. However, out of the 14,000 housekeeping households, 4,183 fail to report household income (29.9%). For regression imputation purposes, household income, by assumption, is a linear function of geographic region, urbanization, race, and household size:[4]

$$
\begin{aligned}
\mathit{INCB4TAX} = {} & 8{,}871.496 + 1{,}670.419 * \mathit{R}1 \\
& (207.641) \qquad (188.731) \\
& + 1{,}810.951 * \mathit{R}2 + 1{,}716.829 * \mathit{R}3 \\
& \quad (187.283) \qquad\qquad (213.063) \\
& - 1{,}667.75 * \mathit{URB}1 \\
& \quad (181.484) \\
& - 2{,}224.44 * \mathit{URB}2 \\
& \quad (170.867) \\
& - 4{,}317.12 * \mathit{RACHH} \\
& \quad (208.808) \\
& + 1{,}490.418 * \mathit{HSIZE} \\
& \quad (42.482)
\end{aligned}
$$

$$R^2 = .1832 \quad F = 314.31 \quad P\text{-VALUE} = .0001.$$

Descriptive statistics of the dependent and regressor variables used in the analysis for both the truncated sample of 9,817 observations and the

[4] A sample selectivity correction was examined in this regression imputation procedure. However, the estimated coefficient associated with the inverse of Mill's ratio (λ_h) was not significantly different from zero at any reasonable level of significance. Consequently, the estimated coefficients used in this regression imputation are free from statistical bias.

Table 4. Parameter Estimates, Standard Errors, Elasticities, and Coefficients of Determination for Alternative Procedures

	Deletion of Missing Values[a]				Regression Imputation[b]				Heckman Procedure[a]		
	OLS	R^2	WLS	R^2	OLS	R^2	WLS	R^2	OLS	R^2	λ_h
Beverages	1.0198[c] (.0713)[d] .3181[e]	.0946	.6629[c] (.0618)[d] .2067[e]	.0968	.9694[c] (.0735)[d] .2942[e]	.0891	.6150[c] (.0598)[d] .1866[e]	.0928	.9772[c] (.0714)[d] .3048[e]	.0988	−42.7812 (6.3252)
Fats & oils	.1090 (.0161) .0772	.2553	.1059 (.0152) .0750	.2638	.1072 (.0162) .0742	.2525	.1059 (.0148) .0732	.2596	.1066 (.0162) .0754	.2556	−2.4872 (1.4350)
Fruits	.4816 (.0442) .1393	.1780	.4679 (.0414) .1353	.1827	.4828 (.0454) .1346	.1668	.4817 (.0410) .1343	.1728	.4706 (.0444) .1361	.1786	−11.0362 (3.9304)
Grains	.1827 (.0532) .0322	.4265	.1886 (.0490) .0333	.4350	.1531 (.0522) .0267	.4277	.1580 (.0464) .0276	.4345	.2113 (.0533) .0373	.4287	28.8060 (4.7236)
Meat & meat alternates	2.3727 (.1535) .1346	.3986	2.0948 (.1434) .1188	.4086	2.2971 (.1551) .1272	.3835	2.0125 (.1393) .1114	.3936	2.3019 (.1539) .1306	.4003	−71.2235 (13.6263)
Milk products	.3899 (.0553) .0679	.4133	.3594 (.0514) .0625	.4190	.3484 (.0559) .0593	.4076	.3198 (.0500) .0544	.4122	.4341 (.0553) .0756	.4181	44.4876 (4.9008)
Miscellaneous foods	.2694 (.0366) .1301	.0779	.2548 (.0341) .1230	.0823	.2522 (.0360) .1186	.0778	.2420 (.0327) .1138	.0818	.2515 (.0367) .1214	.0808	17.9950 (3.2489)
Sugars & sweets	.0357 (.0247) .0298	.1466	.0293 (.0223) .0244	.1548	.0422 (.0235) .0345	.1531	.0355 (.0203) .0290	.1609	.0446 (.0248) .0372	.1480	8.9478 (2.2013)
Vegetables	.6085 (.0523) .1106	.2816	.5920 (.0503) .1076	.2867	.5914 (.0524) .1050	.2721	.5807 (.0492) .1032	.2773	.5766 (.0524) .1048	.2851	−32.1581 (4.6409)
Food at home	5.4697 (.2941) .1192	.5305	4.9726 (.2764) .1084	.5374	5.2442 (.2994) .1117	.5159	4.7681 (.2716) .1015	.5230	5.3748 (.2951) .1171	.5311	−95.4400 (26.1204)

[a] Sample of 9,817 observations.
[b] Sample of 14,000 observations.
[c] Parameter estimate.
[d] Standard error.
[e] Elasticity calculated at the sample means.

sample of 14,000 observations are exhibited in table 3. The means and standard deviations of the respective variables are strikingly similar for the two respective samples. In regard to the household income variable, *INCB4TAX*, the mean for the sample of 14,000 observations exceeds the truncated sample, but the standard deviation is considerably less for the sample of 14,000 observations. Less variability in household income for the original sample is directly attributable to regression imputation for the missing values.

The Engel functions deal not only with aggregate food at home but also nine food groups: (*a*) beverages, (*b*) fats and oils, (*c*) fruits, (*d*) grains, (*e*) meat and meat alternates, (*f*) milk products, (*g*) sugars and sweets, (*h*) vegetables, and (*i*) miscellaneous foods. Households not recording purchases during the specified period but having otherwise complete records are included in the sample. Sample observations with zero expenditure levels are retained to portray adequately the full range of observed behavior. Nevertheless, the percentage of zero observations for the various food groups, with the exception of beverages, is less than 10%. For beverages, this percentage is roughly 17%. For the most part, the estimation of the Engel relationships should not be greatly influenced by zero expenditure levels.

Parameter estimates, associated standard errors, elasticities, and coefficients of determination for the various procedures to handle the missing household income problem are exhibited in table 4. The detection and correction for heteroscedasticity are made via the use of the Park-Glejser procedure and weighted least squares. Correcting for heteroscedasticity in the Heckman procedure is accomplished using the technique developed by Heckman (pp.

480–83). However, this technique breaks down for all but two food groups.[5] Consequently, for the Heckman procedure only the OLS results are presented.

Except for sugars and sweets, the income coefficient estimates are statistically greater than zero at the .10 level of significance. The notable feature, with the exception of beverages, is the similarity of the parameter estimates, standard errors, and elasticities for household income as well as the goodness-of-fit statistics for the respective procedures. Although not reported because of space limitations, this result generally holds for the remaining coefficient estimates of the Engel functions as well. As expected, the estimates of the standard errors diminish for the WLS procedure *vis-à-vis* the OLS procedure. The similarity of the respective parameter estimates across the various procedures may be attributable to the similarity of the samples (see table 3) as well as the size of the samples.

Despite the similarity of the respective parameter estimates and elasticities, the Heckman procedure is statistically superior to the procedure of item deletion. The estimates of the parameter λ_h, derived from the previously discussed probit model specification (table 2), are statistically different from zero in all cases. This result suggests the existence of sample selection bias. Consequently, deleting missing income observations from the analysis is inappropriate from a statistical point of view. However, no other conclusion with regard to superiority of procedure can be drawn on the basis of this empirical evidence. Traditional regression imputation seems to compare favorably with the Heckman procedure. Even the correction of heteroscedasticity, except for beverages, seemingly has little effect on the magnitude of the parameter estimates and the corresponding elasticities.

The inability of this study to distinguish among procedures to handle the missing value problem pinpoints the need for monte carlo simulation. At present, the optimal solution for dealing with the problem of missing observations remains a mystery and thus open to empiricism. Given the importance of the income variable in demand analysis, additional work in this area warrants attention.

[*Received September 1985; final revision received January 1986.*]

[5] This GLS technique requires the estimation of a parameter ρ^2 (Heckman, p. 480). The procedure breaks down when this parameter estimate lies outside the unit interval. Maddala discusses a method to circumvent this problem—the use of the Amemiya estimator. However, because the Amemiya estimator is very cumbersome to apply, it is not used in this study.

References

Aitchinson, J., and J. A. C. Brown. "A Synthesis of Engel Curve Theory." *Rev. Econ. Stud.* 22(1954):35–46.

Amemiya, T. "Qualitative Response Models: A Survey." *J. Econ. Lit.* 19(1981):1483–1536.

Atkinson, A. B., and J. Micklewright. "On the Reliability of Income Data in the Family Expenditure Survey 1970–1977." *J. Royal Statist. Soc.* 146(1983):33–61.

Brown, J. A. C., and A. Deaton. "Surveys in Applied Economics: Models of Consumer Behavior." *Econ. J.* 82(1972):1145–1236.

Buse, R. C., and L. E. Salathe. "Adult Equivalent Scales: An Alternative Approach." *Amer. J. Agr. Econ.* 60(1978):460–68.

Chapman, D. W. "A Survey of Nonresponse Imputation Procedures." *Proceed. Amer. Statist. Assoc. Soc. Statist. Sect.,* 1976, pp. 245–51.

Cox, B. G. "Imputation Procedures to Replace Missing Responses to Data Items." Research Triangle Institute, Research Triangle Park NC, April 1981.

Dagenais, M. G. "The Use of Incomplete Observations in Multiple Regression Analysis." *J. Econometrics* 1(1973):317–28.

Ferber, R. "Consumer Economics, A Survey." *J. Econ. Lit.* 11(1973):1303–42.

Glaser, M. "Linear Regression Analysis with Missing Observations among the Independent Variables." *J. Amer. Statist. Assoc.* 59(1964):834–44.

Goldberger, A. *Econometric Theory.* New York: John Wiley & Sons, 1964.

Heckman, J. J. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimation for Such Models." *Ann. Econ. and Soc. Measure.* 5(1976): 475–92.

Hunter, L., and D. West. "Merging the NFCS and SIE: Methods and Findings." Paper presented to the S-165 Technical Committee on Household Food Consumption, Atlanta GA, Oct. 1982.

Maddala, G. S. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press, 1983.

Pindyck, R. S., and D. L. Rubinfeld. *Econometric Models and Economic Forecasts.* New York: McGraw-Hill Book Co., 1981.

Prais, S. J., and H. S. Houthakker. *The Analysis of Family Budgets.* Cambridge: Cambridge University Press, 1955.

Price, D. W. "Unit Equivalent Scales for Specific Food Commodities." *Amer. J. Agr. Econ.* 52(1970):224–33.