IMPUTING INCOME FOR AN N-PERSON CONSUMER UNIT

Nanak Chand and Charles H. Alexander, U.S. Bureau of the Census
Nanak Chand, U.S. Bureau of the Census, Washington, DC 20233

KEY WORDS: Non-response, stochastic regression, multiple imputation

I. Introduction

The Consumer Expenditure Survey (CE) collects information on expenditure and income from persons fourteen years and older in a consumer unit (CU). Each person in a CU may receive income from one or more sources. The common sources are wage and salary, self-employment, farm, social security, interest, and dividends.

The problem is to impute the missing value of income for any person who receives income but has missing amounts. Earlier work on this project (Crawford (1990) and Paulin and Sweet (1993)) modelled the relationship of income with other variables including expenditure, but did not describe how the model would be used to yield imputed values. The problem is complicated because a given CU may have several earners, each of them could have one or more sources of income.

One important use of income variables is to study the correlation of income with expenditure and other related variables. The imputed income values within any CU are to be consistent with one another and with the characteristics of the person or the CU.

These considerations imply that the imputation of mean values may not be satisfactory. Herzog and Rubin (1983), and Little and Rubin (1987) have suggested a stochastic regression method, where a missing value is replaced by the sum of a regression imputation and a residual drawn to reflect uncertainty in the predicted value.

This paper develops a stochastic regression method for imputing missing income for an N-person CU with several sources of income. The non-response is assumed to be ignorable as defined in Little and Rubin (1987). The procedure consists of generating random variables with replacement to be used as imputed values. The imputation is performed at the person level and takes into account the variability of the observed values of the variables. The solution preserves the observed relationships between the CU members and the sources of income.

The users of the CE data would apply complete-data methods to the imputed income variables. In order to reflect reduced sample size in the resulting standard errors, we are proposing multiple imputation (Rubin (1987)) with the method developed below.

II. Predicted and Imputed Values

The problem is to impute n income variables $Y_1, ..., Y_n$ with missing observed values. We assume that for a CU with any missing value, all the values are missing. For CE, there are very few instances of partial income non-response for a given CU. The Y_i are interrelated since they may represent the same source of income for different CU members or different sources of income for a person or a CU.

It is assumed that a suitable multiple regression model such as described in Martin, Little, Samuel and Triest (1986) is used to predict Y_i . A model for general patterns of missing data for varying CU size and number of sources is described in Section VII.

The following statistics are derived from the model:

m_i = Best prediction of the dependent variable Y_i ,

σ_i = Estimated standard deviation of $Y_i - m_i$, $i=1, ..., n$, and

ρ_{ij} = Estimated correlation coefficient between $Y_i - m_i$ and $Y_j - m_j$, $(i,j) = 1, ..., n$.

The imputed values U_i of Y_p, $i = 1, ..., n$, are derived as described in Sections III and IV.

III. Derivation of Imputed Values

The imputed values are based on the following theorem.

Theorem: let $Z_1, ..., Z_n$ be independent standard normal random variables, and let

$$V_j = \sum_{k=1}^{j-1} a_{jk} Z_k + Z_j + ... + Z_n$$

$$V_1 = \sum_{k=1}^{n} Z_k,$$

where $(a_{ji}, i = 1, ..., j-1, j = 2, ..., n)$ are n(n-1)/2 coefficients given by n(n-1)/2 equations

$$G_{ij} = 0,$$

with

$$G_{ij} = H_{ij} - \rho_{ij},$$

$$H_{ij} = [P(ij) + Q(ij) + n - j + 1]/S(i)S(j),$$

$$P(ij) = \sum_{k=1}^{i-1} a_{ik} a_{jk} \quad (P(1j) = 0)$$

$$Q(ij) = \sum_{k=i}^{j-1} a_{jk},$$

$$S(\eta) = S_n(\eta) = [R(\eta) + n - \eta + 1]^{1/2} \quad (S(1) = \sqrt{n}),$$

and

$$R(\eta) = \sum_{k=1}^{\eta-1} a_{\eta k}^2, \; \eta = i, j.$$

Then the random variables

$$U_k = m_k + (\sigma_k / S(k)) V_k, \; k = 1, ..., n$$

are distributed as

$$N(m_k, \sigma_k^2)$$

with

$$\rho_{ij} = Corr \; (U_i, U_j), \; i = 1, ..., j-1, \; j = 2, ..., n.$$

Proof: by Induction

IV. The Imputation Procedure Given Estimates of the Parameters

The procedure is to draw n standard normal variables

$$Z_1, ..., Z_n$$

and to transfer them to give imputed values $U_1, ..., U_n$ as defined in Section III.

V. Alternative Methods

Choleski factorization decomposes a variance-covariance matrix into a product of a triangular matrix and its transpose, and thus provides an alternative to the above method. Due to its triangular nature, however, this approach would impute y_1 as a function of Z_1, Y_2 as a function of (Z_1, Z_2), and so on.

We prefer the method given in Sections III-IV since it is most general in the sense that it imputes each Y_k, $k = 1, ..., n$ as a function of $(Z_1, ..., Z_n)$.

Intermediate methods would base imputation of Y_k on $(Z_1, \ldots, Z_l), k < l < n,\ k = 1, \ldots, n.$

VI. Solving for Coefficients (a_{jl})

The coefficients (a_{jl}) in Section III are given by n-1 sets of equations. The (j-1)th set consists of j-1 equations and is given by

$$\rho_{ij} = H_{ij},\ i = ,\ \ldots,\ j-1, j = 2, ., ., n,$$

Dividing the ith equation in the (j-1)th set by the first equation in the set, and defining, $L = L(ij) = \rho_{ij} S(i)/\sqrt{n}\ \rho_{1j}$, L being independent of $(a_{jk},\ k = 1, \ldots, j-1)$, we have

$$E_{ij} = 0,\ where$$

$$E_{ij} = P(ij) + Q(ij) - LQ(1j) + (1-L)(n-j+1) =$$

$$\sum_{k=1}^{i-1} a_{jk}(a_{ik} - L) + \sum_{k=1}^{j-1} a_{jk}(1-L) + (1-L)\ (n-j+1),$$

$$i = 2, \ldots, j-1,\ j = 2, \ldots, n.$$

For fixed $(a_{ik},\ i = 1, \ldots, j-1,\ k = 1, \ldots, j-1)$ the above process results in j-2 linear equations in j-1 unknown coefficients $(a_{jk},\ k = 1, \ldots, j-1)$.

Each of $(a_{jk},\ k = 2, \ldots, j-1)$ may thus be expressed in terms of a_{j1}. Substituting these expressions in the first equation of the set $(G_{ij} = 0)$, gives a_{j1}, and hence $(a_{jk},\ k = 1, \ldots, j-1)$.

VII. A Model for General Patterns of Missing Data Due to Varying CU Size and Number of Sources

Let Y_i be the outcome variable, and $\{X_{ij}, j=1, \ldots m\}$ be the corresponding independent variables for predicting Y_i for the ith unit (ith CU member or ith source of income), i=1,..., n. A proposed model for the ith unit is:

$$E\,Y_i = \alpha_i + \sum_{j=1}^{m} \beta_{ij} X_{ij} + \sum_{k=1}^{i-1} \gamma_{ik} \hat{Y}_k \text{ ,where } \hat{Y}_k \text{ is}$$

the predicted value of Y_k for the kth unit, k=1,...,i-1, and $\{\beta_{ij}$, j=1,...,m, γ_{ik}, k=1,...,i-1$\}$ are unknown parameters. The errors are assumed to be independently and normally distributed with equal variances. We are developing procedures to handle non-normal errors. In addition, we are estimating the effect of the second set of terms in the model.

The independent variables used for modelling wage and salary are listed below:

Dependent Variable Y : Log (Salary)

Independent Variables:

X1	:	Age of member
X2	:	Squared age
X3	:	Log (Hours worked per week)
X4	:	Log (Weeks worked per year)
X5	:	Grades completed (1-12, 13-16, 17-20)
X6	:	Number of vehicles in the family

X7 : Number of rooms in the CU

X8 : Categorical variable (CV) indicating whether the CU had income from interest

X9 : CV indicating if the member has a job

X10 : CV indicating if the member is a fulltime college student

X11 : CV having values of 1 for male and 2 for female

X12 : CV having a value 1 if the employer contributes to member's pension plan, 2 otherwise

X13 : CV indicating whether the member put money in a retirement account

X14 : CV indicating if the member is working full time for full year, part time for full year, full time for part of the year, or part time for part of the year

X15 : CV indicating whether the CU resides in an apartment, a mobile home, or a college dormitory

X16 : CV indicating whether the CU received food stamps

X17 : CV indicating whether the member received supplemental security income during the past year

VIII. Illustration

The above procedure of imputing income variables is illustrated in the following example by imputing wage and salary, three times each, for 3 person CUs, contained in the 1988-1990 Consumer Expenditure data resulting from the second interviews. The total number of observations is 1,509.

Example: Imputing Wage and Salary for a Three-Person CU

No. of Observations = 1509 for first CU member
No. of Obersvations = 926 for second CU member
No. of Observations = 234 for third CU member

σ_1 = 0.660557

σ_2 = 0.687700

σ_3 = 0.660235

ρ_{12} = 0.183640

ρ_{13} = 0.150330

ρ_{23} = 0.0855890

OBS	Predicted Values of Log (Salary)		
	Member1	Member2	Member3
1	10.7910	8.8332	6.2057
2	10.1827	9.3249	7.8510
3	10.2326	8.8934	8.4631
4	10.5771	9.7275	9.6375
5	10.5572	10.1317	7.0451

OBS	First Imputation Normal (0,1) Variables		
	Z1	Z2	Z3
1	0.68061	0.54360	-0.58221
2	0.34520	0.69708	-1.58979
3	-0.91822	0.68287	1.16646
4	-0.65000	0.49918	0.67965
5	-1.24836	0.13455	-1.22924

OBS	Imputed Values		
	U1	U2	U3
1	11.035841	8.493688	5.672579
2	9.973895	8.848076	6.854057
3	10.587700	9.978480	8.573640
4	10.778781	10.450160	9.650697
5	9.663625	10.347790	6.528663

OBS	Second Imputation Normal (0,1) Variables		
	Z1	Z2	Z3
1	-1.35567	-0.45695	0.04560
2	-0.05624	-0.14934	0.91303
3	0.61550	-0.47717	-0.44149
4	1.66517	-0.82197	-0.42728
5	0.43744	-1.14995	0.40742

OBS	Imputed Values		
	U1	U2	U3
1	10.117107	9.339107	6.501715
2	10.452502	9.618286	8.285903
3	10.116983	8.278038	8.531265
4	10.735720	8.493961	9.868667
5	10.440847	9.663086	7.805973

OBS	Third Imputation Normal (0,1) Variables		
	Z1	Z2	Z3
1	0.57517	-1.42054	0.02276
2	-0.10603	-1.60708	-2.22514
3	-0.16596	-1.76976	0.64415
4	-0.89746	-0.22399	1.33022
5	-1.28959	-0.17863	-1.11929

OBS	Imputed Values		
	U1	U2	U3
1	10.477279	8.069997	6.958499
2	8.680760	8.038667	7.849769
3	9.740031	8.580187	9.661934
4	10.656719	10.543375	10.295198
5	9.570395	10.296636	6.739547

IX. Future Research

We are planning to use multiple imputation in conjunction with the method described in this paper. However, practical considerations may suggest utilizing one of the variations of the full Bayesian multiple imputation.

These variations pertain to creating multiply-imputed data sets by randomly selecting the residual term only or by randomly selecting the mean value and the residual term only.

A different topic relates to determining hierarchy of income variables for the model of Section VII.

REFERENCES

Crawford, S. Several Internal Memos from the Bureau of Labor Statistics, 1989-1990.

Herzog, T. N., and Rubin, D. B. (1983). Using Multiple Imputation to Handle Non-response in Sample Surveys, in Incomplete Data in Sample Surveys, Vol. II. New York: Academic Press.

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

Martin, D., Little, R. J. A., Samuel, M. E., and Triest, R. K. L. (1986). Alternative Methods for CPS Income Imputation, Journal of the American Statistical Association, 81. 29-41.

Paulin, G. and Sweet, E. (1993). Modelling Income in the U.S. Consumer Expenditure Survey. Paper presented at the Annual Meeting of the American Statistical Association.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Sample Surveys. New York: Wiley.