ASYMPTOTIC PROPERTIES OF IMPUTED VARIABLES IN THE CONSUMER EXPENDITURE SURVEY

Nanak Chand , Charles H. Alexander, U.S. Bureau of the Census
Nanak Chand, U.S. Bureau of the Census, Washington, D.C. 20233

Key words: Item nonresponse, Bayesian imputation

1. Introduction

The Consumer Expenditure Survey (U.S. Department of Labor (1993)) collects information on expenditures and income in a consumer unit. Nearly fifteen percent of the sampled units have item nonresponse on the income questions even though they have enough expenditure data to count as "interviewed units". An important use of income variables is to study the correlation of income with expenditures and other related variables. Thus the objective is to impute the missing value of income for any person with missing amounts, so that the imputed variables are consistent with one another and with the characteristics of the person and the consumer unit.

The imputation process models the relationship of income with other variables based on data from units with complete response, and uses these models to produce imputed values. The general goal is that the distribution of the imputed income values conditional on other variables describing the household should be realistic. To achieve this goal, a model is adopted for the relationship of the variables, and the parameters of this model are estimated from the observed units that have complete data. The imputed income values for any household are then generated from the distribution implied by the model with the estimated parameters. This paper explores simpler alternatives to produce such imputations and establishes their asymptotic equivalence. A simulation study demonstrates this equivalence at the various sample size levels.

2. The Model

The income vector $\underline{y} = (y_1, \ldots y_n)$ of observable random variables representing the data from units with complete data, follows the regression model

$$\underline{y} = X\underline{\beta} + \underline{e}$$

where X is an nxq (n>q) matrix of known transformed values, $\underline{\beta}$ is a qx1 vector of unknown regression coefficients, $\underline{e}$ is an nx1 vector of random errors, and the following assumptions from the econometric literature (Fomby, Hill and Johnson (1984), and Vinod and Ullah (1981)) hold:

A_1: X is non-stochastic,

A_2: $rank(X^TX) = q$,

A_3: $\lim_{n\to\infty} \frac{X^TX}{n} = Q$,

where Q is a finite and nonsingular matrix, and

A_4: $\underline{e}$ is multivariate normal with mean vector $\underline{0}$ and covariance matrix $\sigma^2 I$.

The least squares estimator $\underline{b}$ of $\underline{\beta}$ is the maximum likelihood estimator under A_4, and is given by

$$\underline{b} = \Sigma X^T\underline{y},$$

where

$$\Sigma = (X^TX)^{-1}$$

This is an unbiased estimator of $\underline{\beta}$ with the covariance matrix given by

$$v(\underline{b}) = \sigma^2\Sigma$$

We can write

$$\hat{e} = \underline{y} - X\underline{b}$$

as an estimator of $\hat{e}$, with

$$s^2 = \frac{\hat{e}^T\hat{e}}{n-q}$$

being a consistent and unbiased estimator of σ^2.

For a missing variable y_0 with the corresponding observed (qx1) vector $\underline{x}_0$, the mean predicted value is given by

$$\hat{m}_0 = \underline{x}_0^T\underline{b}$$

where

$$m_0 = \underline{x}_0^T\underline{\beta}.$$

3. Imputation of Missing Values

We consider three alternative stochastic imputations of y_0. The multiple imputation task is accomplished by independently drawing additional random variables for each of these three cases similarly as for the single imputation given below.

The first type of imputation is obtained by drawing a N (0, 1) variate z_0 and taking imputed variable as

$$U^{(1)} = \hat{m}_0 + s z_0$$

This imputation is consistent with the random nature of the error term and its assumed distribution. This is a non-Bayesian imputation since it is based on the estimators $\underline{b}$ and s^2 as given by the model.

The second type of imputation is derived by taking the conventional prior distribution of $(\underline{\beta}, \sigma^2)$ as

$$p(\underline{\beta}, \sigma^2) \propto \sigma^{-2}.$$

Then the marginal posterior distribution of σ^2 given $\underline{y}$ is a reciprocal χ^2 given by

$$\sigma^2 \sim \frac{\nu s^2}{\chi^2_\nu}$$

where

$$\nu = n - q.$$

We obtain the second type of imputation, which is partial - Bayesian since it is based on the above marginal posterior distribution of σ^2 by drawing a χ^2_ν random variable and imputing y_0 by the variable

$$U^{(2)} = \hat{m}_0 + \frac{\sqrt{\nu} s z_0}{g},$$

where g is the square root of the χ^2_ν variable.

The third type of imputation, termed as fully Bayesian to be based on the posterior distributions of $\underline{\beta}$ and σ^2, is derived as follows:

Under the above joint prior distribution of (β, σ^2), we have, a posteriori,

$$\underline{\beta} \sim N(\underline{b}, \sigma^2 \Sigma)$$

and

$$\sigma^2 \sim \frac{\nu s^2}{\chi^2_\nu}.$$

This imputation is thus obtained by drawing additional q independent N (0, 1) variables giving a qx1 vector

$$\underline{z} = (z_1, \ldots z_q),$$

and by imputing y_0, under the assumption of positive definiteness of the matrix Σ, by the variable

$$U^{(3)} = \hat{m}_0 + \frac{\sqrt{\nu} s z_0}{g} + \frac{\underline{x}_0^T \Sigma^{1/2} \underline{z} \sqrt{\nu} s}{g},$$

where $\Sigma^{1/2}$ is a square root of Σ.

4. Expected Values, Variances, and Distributions of the Imputed Variables

4a. Small Sample Results

The expression for $U^{(3)}$ can be simplified to

$$U^{(3)} = \hat{m}_0 + \frac{\sqrt{\nu} s h \bar{z}}{g},$$

where

$$\bar{z} = \frac{z_0 + \underline{x}_0^T \Sigma^{1/2} \underline{z}}{h},$$

with

$$h = h(\underline{x}_0, \Sigma) = \sqrt{1 + \underline{x}_0^T \Sigma \underline{x}_0}.$$

The positive definiteness of Σ implies that h exists and is positive. Also, $\bar{z}$ has a normal distribution with mean 0 and variance 1.

Since the ratio of a N (0, 1) variate to a $\sqrt{\chi^2_\nu / \nu}$ variate is distributed as a t variable with ν degrees of freedom, we can write

$$U^{(1)} = \hat{m}_0 + s z_0,$$

$$U^{(2)} = \hat{m}_0 + s t_\nu,$$

and

$$U^{(3)} = \hat{m}_0 + s h t_\nu,$$

where t_ν is a t variate with ν degrees of freedom. In addition, for i = 1, ..., 3, we have,

$$E(U^{(i)}) = E[E(U^{(i)} \mid b, s)]$$

$$= E(\hat{m}_0)$$

$$= m_0$$

and letting

$$V^{(i)} = U^{(i)} - \hat{m}_0,$$

it follows that

$$v(V^{(1)}) = E[E(z_0^2 s^2 | s^2)]$$

$$= E(s^2) = \sigma^2 ,$$

$$v(V^{(2)}) = E[E(s^2 t_\nu^2 | s^2)]$$

$$= \frac{\nu}{\nu - 2} E(s^2) = \frac{\nu}{\nu - 2} \sigma^2 ,$$

and

$$v(V^{(3)}) = E[E(h^2 t_\nu^2 s^2 | s^2)]$$

$$= \frac{\nu}{\nu - 2} h^2 \sigma^2.$$

Also for i = 1, ..., 3, we have,

$$v(U^{(i)}) - v(V^{(i)}) = E[\hat{m}_0 - m_0]^2$$

$$= \sigma^2 \underline{x_0}^T \Sigma x_0 = \sigma^2 (h^2 - 1).$$

4b. Large Sample Results

Since the matrix Q given by

$$Q = \lim_{n \to \infty} \frac{X^T X}{n}$$

is finite and nonsingular, the matrix

$$Q^{-1} = \lim_{n \to \infty} (X^T X / n)^{-1}$$

is also finite and nonsingular.

Therefore the sequence of elements within $n(X^T X)^{-1}$ is bounded. Thus there exists a real number C such that

$$\underline{x}_0^T n (X^T X)^{-1} \underline{x}_0$$

$$= \underline{x}_0^T \Sigma \underline{x}_0 / n^{-1} \leq C$$

for all n, or

$$h^2 - 1 = \underline{x}_0^T \Sigma \underline{x}_0 = O(\frac{1}{n}),$$

that is, $\underline{x}_0^T \Sigma \underline{x}_0$ is at most of order 1/n.

The following results follow:

$$v(U^{(1)}) = \sigma^2 h^2$$

$$= \sigma^2 [1 + O(\frac{1}{n})]$$

$$v(U^{(2)}) = [\frac{\nu}{\nu - 2} + h^2] \sigma^2$$

$$= \sigma^2 [(O(\frac{1}{\nu}) + 1 + O(\frac{1}{n})]$$

$$= \sigma^2 [1 + O(\frac{1}{n})]$$

$$v(U^{(3)}) = \frac{\nu}{\nu - 2} h^2 \sigma^2 + \sigma^2 (h^2 - 1)$$

$$= \sigma^2 [(1 + O(\frac{1}{\nu}))(1 + O(\frac{1}{n})) + O(\frac{1}{n})]$$

$$= \sigma^2 [1 + O(\frac{1}{n})]$$

Also, since the t_ν distribution converges to the normal distribution for large ν, the following theorem follows:

Theorem:

The asymptotic mean (ASY.E) and the asymptotic variance (ASY.v) of each of the imputed variables $U^{(i)}$ are equal and are given by

$$ASY.E(U^{(i)}) = m_0,$$

$$ASY.v(U^{(i)}) = \sigma^2,$$

and the asymptotic distribution (ASY.D) of the random variable

$$W^{(i)} = \frac{U^{(i)} - \hat{m}_0}{s},$$

is given by

$$ASY.D(W_i) \sim N(0, 1),\ i = 1, \ldots, 3.$$

5. Inferences Based on the Complete Data Sets

The complete data set consists of proportion w of known variables with mean $\bar{V}$, $\bar{U}_j^{(i)}$ being the corresponding mean of the remaining proportion (1-w) of k imputed variables resulting from applying the j_{th} imputation of type i with associated variance σ_i^2, i = 1, ..., 3; j = 1, ..., m.

An estimate of the population mean M is

$$M_{ij} = w\bar{V} + (1-w)\bar{U}_j^{(i)}$$

The average estimate over the m imputations is

$$\bar{M}_i = w\bar{V} + (1-w)\bar{U}^{(i)},$$

where

$$\bar{U}^{(i)} = (1/m)\sum_{j=1}^{m} \bar{U}_j^{(i)},$$

and the average of the corresponding variances is

$$\bar{W}_i = \frac{(1-w)}{k}[w\sigma^2 + (1-w)\sigma_i^2]$$

The expected value of the between imputation variance for the i_{th} type of imputation is

$$B_i = \frac{1}{m-1}\sum_{j=1}^{m} E(M_{ij} - \bar{M}_i)^2$$
$$= \frac{(1-w)^2}{m-1}\sum_{j=1}^{m} E(\bar{U}_j^{(i)} - \bar{U}^{(i)})^2$$

The total variance of $M - \bar{M}_i$ is thus given by

$$\delta_i^2 = \bar{W}_i + \frac{m+1}{m}B_i$$

Let

$$CI = CI[M - \bar{M}_i]$$

denote the 100 (1- α) % confidence interval for $M - \bar{M}_i$. Then

$$CI = [-\delta_i t_d(\alpha/2),\ +\delta_i t_d(\alpha/2)],$$

where $t_d(\alpha/2)$ is the upper $100(\alpha/2)$ percentage point of the t distribution with

$$d = \frac{(m-1)\delta_i^4}{(\delta_i^2 - \bar{W}_i)^2}$$

degrees of freedom and

$$\frac{(\delta_i^2 - \bar{W}_i)}{\bar{W}_i} = \frac{m+1}{m}\frac{B_i}{\bar{W}_i}$$

is the relative increase in variance due to nonresponse.

Expressing $\bar{U}_j^{(i)}$ in terms of mean of the $\hat{m}_0$ values and the randomly selected variables as in the earlier sections, we have, for i = 2, 3,

$$\lim_{n\to\infty} E(\bar{M}_i - \bar{M}_1)^2$$
$$= \lim_{n\to\infty} v(\bar{M}_i - \bar{M}_1) = 0$$

The Chebyshev inequality now implies that

$$plim_{n\to\infty} E(\bar{M}_i - \bar{M}_1) = 0$$

In addition, the expression for B_i simplifies to

$$B_i = \frac{(1-w)^2}{k}\sigma^2[1 + O(\frac{1}{n})]$$

Also since,

$$\bar{W}_i = \frac{(1-w)}{k}\sigma^2[1 + O(\frac{1}{n})] ,$$

we have, ignoring terms of order $\frac{1}{n}$,

$$CI = [-\delta t_d(\alpha/2), + \delta t_d(\alpha/2)] ,$$

where

$$\delta^2 = \frac{(1-w)}{k}\sigma^2 [1 + \frac{m+1}{m}(1-w)]$$

and

$$d = (m-1)[1 + \frac{m}{m+1}(1-w)^{-1}]^2 ,$$

$$i = 1, ..., 3.$$

6. Simulations

The simulation study is based on the 1988-1990 consumer expenditure data (U.S. Department of Commerce (1993)) resulting from the second interviews. The variable under consideration is the logarithm of the salary of the reference person in a consumer unit, who according to the survey respondent, is the one who owns or rents the unit. The number n of reference persons in this data set with observed salary and other related variables is 7,686.

For comparing variances of $U^{(i)}$, i = 1, ..., 3, we consider the magnitudes of the following two expressions:

$$D_{12} = \frac{v(V^{(2)}) - v(V^{(1)})}{v(V^{(1)})}$$

$$= \frac{2}{\nu - 2},$$

and

$$D_{23} = \frac{v(V^{(3)}) - v(V^{(2)})}{v(V^{(2)})}$$

$$= \underline{x}_0^T \Sigma \underline{x}_0$$

$D_{ij} = D_{ij}(V)$ is the proportional difference in variance between $V^{(j)}$ and $V^{(i)}$, and since

$$D_{ij}(U) \le D_{ij}(V)$$

$D_{ij}(V)$ gives the upper bound of the relative increment in variance in selecting $U^{(j)}$ over $U^{(i)}$.

The simulations are performed by taking 1,000 samples each of size 500 and 1, 000 to 5,000 in increments of 1, 000 of the following set of design variables from the above data.

The design matrix consists of the following five columns:

X_1: Intercept

X_2: Age

X_3: Logarithm of the number of hours worked per week

X_4: Logarithm of the weeks of work during the year

X_5: Highest grade completed in school

Table B provides relative mean difference values D_{12} and D_{23} for the 1,000 samples for each of the selected sample size levels for the set of $\underline{x}_0^j = (x_{01}^j, ..., x_{05}^j)$ vectors, j = 1, ..., 10, given in Table A.

The table shows small relative differences in variances between $U^{(1)}$ and $U^{(2)}$ and extremely small such differences between those of $U^{(2)}$ and $U^{(3)}$, as indicated by the respective upper bounds. The desirability of the choice of $U^{(2)}$ is obvious considering both the magnitude of variance and the computational aspects of arriving at the multiply-imputed data sets.

References

[1] Fomby, T.B., Hill, R.C., and Johnson, S.R. (1984). Advanced Econometric Methods. New York: Springer-Verlag.

[2] U.S. Department of Commerce (1993). Internal Data Files on the Consumer Expenditure Survey, 1988 - 1990, Bureau of the Census, Washington, D.C. 20233.

[3] U.S. Department of Labor (1993). Consumer Expenditure Survey, 1990 - 1991, Bulletin 2425, Bureau of the Labor Statistics, Washington, D.C. 20212.

[4] Vinod, H. D., and Ullah, A. (1981). Recent Advances in Regression Methods. New York: Marcel Dekker, Inc.

Table A

$\{x_{oi}^{j}\}$ Values

$x_{01}^{j} = 1$

	Age	Log [Number of Hours Worked per Week]
j	x_{02}^{j}	x_{03}^{j}
1	25	3.80666
2	27	3.73767
3	33	3.80666
4	38	3.82864
5	46	3.68888
6	49	3.91202
7	53	3.68888
8	57	3.80666
9	60	3.61092
10	66	3.68888

Table A (Continued)

	Log [Number of Weeks of Work During the Year]	Highest Grade Completed
j	x_{04}^{j}	x_{05}^{j}
1	3.91202	16
2	3.93183	14
3	3.91202	16
4	3.95124	15
5	3.89182	17
6	3.87120	14
7	3.91202	19
8	3.87120	20
9	3.87120	18
10	3.85015	17

Table B
Upper Bounds on The
Relative Increment in Variance
Number of Simulated Samples: 1,000

Sample Size

	500	1,000	2,000
j/D_{12}	.0041	.0020	.0010
	$10^6 \times D_{23}$		
1	1.0244	.5081	.2527
2	.7337	.3640	.1815
3	.6873	.3414	.1699
4	.4885	.2429	.1211
5	.9703	.4819	.2398
6	.6886	.3404	.1694
7	2.0704	1.0274	.5112
8	2.8979	1.4377	.7156
9	2.2295	1.1058	.5500
10	2.5000	1.2412	.6179

Table B (Continued)

	3,000	4,000	5,000
j/D_{12}	.0007	.0005	.0004
	$10^6 \times D_{23}$		
1	.1683	.1262	.1009
2	.1208	.0906	.0724
3	.1132	.0849	.0679
4	.0807	.0605	.0484
5	.1599	.1198	.0958
6	.1126	.0843	.0675
7	.3408	.2554	.2043
8	.4768	.3574	.2858
9	.3665	.2746	.2197
10	.4116	.3084	.2467