User's Guide to Income Imputation in the CE
December 1, 2006

US Department Of Labor
Bureau of Labor Statistics
Division of Consumer Expenditure Surveys

Bureau of Labor Statistics
Consumer Expenditure Survey

Authors: Geoffrey Paulin, DCES
Jonathan Fisher, DPINR
Sally Reyes-Morales, SMD
Initial Version: 12/13/2005
Last Update: 12/01/2006

Multiple Imputation Manual:
Supplement to 2004 Consumer Expenditure Interview Survey Public Use Microdata Documentation

I. BACKGROUND

The purpose of this manual is to provide instructions to users regarding the proper use of multiply imputed data to draw statistically valid inferences in their works. Therefore, the main portion of this text describes application and usage of multiply imputed data, rather than its production or its statistical properties and derivations. However, for data users who are interested in better understanding them, detailed descriptions of the theoretical underpinnings of this process are documented elsewhere.[1]

A. Introduction and Method Overview.

Starting with the publication of the 2004 data, the Consumer Expenditure Surveys include income data that have been produced using multiple imputation. The purpose of this procedure is to fill in blanks due to nonresponse (i.e., the respondent does not know or refuses to provide a value for a source of income received by the consumer unit or a member therein) in such a way that statistical inferences can be validly drawn from the data. The process preserves the mean of each source of income, and also yields variance estimates that take into account the uncertainty built into the data from the fact that some observations are imputed, rather than reported.

The method used to derive the multiple imputations is regression-based. Essentially, a regression is run to provide coefficients for use in estimating values for missing data points. The coefficients are then "shocked" by adding random noise to each, and missing values are estimated using the shocked coefficients. To each of these estimated values, additional random noise is added, to ensure that consumer units (or members) with identical characteristics (e.g., urban service worker aged 25 to 34) will not receive identical estimates for their income. The resulting values are used to fill in invalid blanks where they occur, while reported values are retained. This process is then repeated four times, so that a total of five imputed values are computed for each missing value. In addition, for the small number of cases in which the respondent does not report receipt of any source of income collected either at the member or consumer unit level, receipt of each source is imputed using logistic regression. In each case where receipt is imputed, the income value is treated as a missing data point, and is imputed using the method described above.

B. Historical Income Data Differences and Guidelines for use of Imputed Data.

Starting with the publication of the 1972-73 data, the Consumer Expenditure Survey introduced the concept of the "complete income reporter." In general, consumer units are defined as complete income reporters if their respondents provide values for at least one of the major sources of income, such as wages and salaries, self-employment income, and Social Security income. However, even complete income reporters may not have provided a full accounting of all sources of income. The first difference, therefore,

[1] Rubin, Donald B. Multiple Imputation for Nonresponse in Surveys (New York: John Wiley and Sons, Inc., 1987).

between the data previously published and those available starting in 2004 is that the imputed data have all invalid missing values filled in, so that estimates using income can be generated for all consumer units, not only for complete income reporters.

In addition, the collected data contain only one observation of each income value for each consumer unit or member for whom a value is reported. The imputed data include five estimates of each observation, plus one additional estimate representing the mean of all five estimates. For example, when examining the collected data for a subset of interest (say, 100 particular consumer units who all report receipt of INTEARNX), there is one column of data identifying the selected consumer units (i.e., 100 observations of NEWID) and one column of data containing the associated income values of interest (i.e., 100 observations of INTEARNX). However, with the imputed data, there are five columns of data (each containing 100 observations), each of which has a different value for income if the original value (INTEARNX) is missing due to an invalid nonresponse, or the same value as the original value, if the original value is provided by the respondent. In addition, there is a sixth column of data (also containing 100 observations) that contains the mean of the five columns of data just described.

A common assumption may be that it does not matter which column is used in data analysis, so it is reasonable to select one randomly and use it to draw inferences. Unfortunately, using one column of data in this way does not adequately capture the uncertainty built into the data by the very nature that some of it has been imputed rather than collected from the respondent. Therefore, at a minimum, variance estimates obtained from using one column of data will be biased downward. Proper variance estimation requires use of the five columns of imputed data. Similarly, proper calculation of the estimated mean requires averaging the estimates from all five columns of data. However, it can be shown that finding the average of the 500 observations (that is, the five columns of imputed data for each of the 100 consumer units selected for examination) yields the same answer as averaging each of the five imputations to get one column of 100 imputed means, and then finding the mean of the 100 observations. (See "Computing Means.") Therefore, the sixth column is included as a convenience for users who are interested only in calculating means. However, it is not recommended that users who want to compute variances, regression parameters, or other statistical results use only the sixth column in their analyses.

C. Variable Names.

Imputed income data appear on both the MEMB and FMLY files. Their names are as follows:

Income variable name: MEMB file	Associated 5 imputed income variables	Mean imputed income variable
SALARYX	SALARYX1 - SALARYX5	SALARYXM = mean (SALARYX1 - SALARYX5)
NONFARMX	NONFARM1 - NONFARM5	NONFARMM = mean(NONFARM1 - NONFARM5)
FARMINCX	FARMINC1 - FARMINC5	FARMINCM = mean(FARMINC1 - FARMINC5)
RRRETIRX	RRRETIR1 - RRRETIR5	RRRETIRM = mean(RRRETIR1 - RRRETIR5)
SOCRRX	SOCRRX1 – SOCRRX5	SOCRRM = mean(SOCRRX1 – SOCRRX5)
SSIX	SSIX1 - SSIX5	SSIXM = mean(SSIX1 - SSIX5)

Income variable name: FMLY file	Associated 5 imputed income variables	Mean imputed income variable
PENSIONX	PENSION1- PENSION5	PENSIONM=mean(PENSION1- PENSION5)
INTEARNX	INTEARN1- INTEARN5	INTEARNM=mean(INTEARN1- INTEARN5)
FININCX	FININCX1- FININCX5	FININCXM=mean(FININCX1- FININCX5)
INCLOSSA	INCLOSA1- INCLOSA5	INCLOSAM=mean(INCLOSA1- INCLOSA5)

INCLOSSB	INCLOSB1- INCLOSB5	INCLOSBM=mean(INCLOSB1- INCLOSB5)
UNEMPLX	UNEMPLX1- UNEMPLX5	UNEMPLXM=mean(UNEMPLX1- UNEMPLX5)
COMPENSX	COMPENS1- COMPENS5	COMPENSM=mean(COMPENS1- COMPENS5)
WELFAREX	WELFARE1- WELFARE5	WELFAREM=mean(WELFARE1- WELFARE5)
CHDOTHX	CHDOTHX1- CHDOTHX5	CHDOTHXM=mean(CHDOTHX1- CHDOTHX5)
ALIOTHX	ALIOTHX1- ALIOTHX5	ALIOTHXM=mean(ALIOTHX1- ALIOTHX5)
OTHRINCX	OTHRINC1- OTHRINC5	OTHRINCM=mean(OTHRINC1- OTHRINC5)
FOODSMPX	FOODSMP1-FOODSMP5	FOODSMPM=mean(FOODSMP1-FOODSMP5)
FINCBTAX*	FINCBTX1- FINCBTX5	FINCBTXM=mean(FINCBTX1- FINCBTX5)
FINCATAX*	FINCATX1- FINCATX5	FINCATXM=mean(FINCATX1- FINCATX5)
FSALARYX*	FSALARY1-FSALARY5	FSALARYM =mean(FSALARY1-FSALARY5)
FNONFRMX*	FNONFRM1-FNONFRM5	FNONFRMM =mean(FNONFRM1-FNONFRM5)
FFRMINCX*	FFRMINC1-FFRMINC5	FFRMINCM =mean(FFRMINC1-FFRMINC5)
FRRETIRX*	FRRETIR1-FRRETIR5	FRRETIRM =mean(FRRETIR1-FRRETIR5)
FSSIX*	FSSIX1-FSSIX5	FSSIXM =mean(FSSIX1-FSSIX5)

* Summary variable created from MEMB file data.

D. Other Related Variables.

Additional variables are also available that are created from, or related to, the imputed income variables. These include INC_RANKn and various descriptor variables (section D2. below), which describe the reason for imputation.

D1. INC_RANKn.

As described in the main documentation, INC_RANK is created using complete income reporters only. They are sorted in ascending order of FINCBTAX, and ranked according to a weighted population rank, so that quintiles and other distributional measures can be obtained.

For the imputed data, INC_RANK1 through INC_RANK5 and INC_RANKM are also created in a similar way. The difference is that they each use all consumer units, instead of complete reporters only, and that they are based on sorts of FINCBTXn. (That is, INC_RANK1 is derived from FINCBTX1.)

D2. Descriptor Variables.

Imputation descriptor variables are coded to describe whether the income variable has undergone multiple imputation, and if so, for what reason. The imputation descriptor variable for each income variable is defined in the following tables.

MEMBER INCOME VARIABLES

Income variable name	**Associated imputation descriptor variable**
SALARYX	SALARYXI
NONFARMX	NONFARMI
FARMINCX	FARMINCI

RRRETIRX	RRRETIRI
SOCRRX	SOCRRXI
SSIX	SSIXI

FMLY INCOME VARIABLES

Income variable name	**Associated imputation descriptor variable**
PENSIONX	PENSIONI
INTEARNX	INTEARNI
FININCX	FININCXI
INCLOSSA	INCLOSAI
INCLOSSB	INCLOSBI
UNEMPLX	UNEMPLXI
COMPENSX	COMPENSI
WELFAREX	WELFAREI
FOODSMPX	FOODSMPI
CHDOTHX	CHDOTHXI
ALIOTHX	ALIOTHXI
OTHRINCX	OTHRINCI

SUMMARY FMLY INCOME VARIABLES*

Summary FMLY income variable	**Associated imputation descriptor variable**
FSALARYX	FSALARYI
FNONFRMX	FNONFRMI
FFRMINCX	FFRMINCI
FRRETIRX	FRRETIRI
FSSIX	FSSIXI
FINCBTXM	FINCBTXI

* These represent the sum of the member-level income variables for each family.

Each descriptor variable has a numeric value three characters in length. There are no blanks or blank codes (such as “A”, “D” or “T”) for descriptor variables. The descriptor variables are defined as follows:

CODE VALUE	**CODE DESCRIPTION**
100*	no multiple imputation – reported income is a valid value, or valid blank
201	multiple imputation due to invalid blank only
301	multiple imputation due to bracketing only
501	multiple imputation due to conversion of a valid blank to an invalid blank (occurs only when initial values for all sources of income—MEMB and FMLY--

	for the consumer unit were valid blanks)

* Note that when no imputation occurs, the assigned code value is 100 at both the individual source level and at the summary level.

Description of code values for Summary FMLY Income imputation descriptor variables

CODE VALUE	CODE DESCRIPTION
100	No imputation. This would be the case only if NONE of the variables that are summed to get the summary variables is imputed.
2nn	Imputation due to invalid blanks only. This would be the case if there are no bracketed responses, and at least one value is imputed because of invalid blanks.
3nn	Imputation due to brackets only. This would be the case if there are no invalid blanks, and there is at least 1 bracketed response
4nn	Imputation due to invalid blanks AND bracketing.
5nn	Imputation due to conversion of valid blanks to invalid blanks. (Occurs only when initial values for all sources of income for the consumer unit and each member are valid blanks.)

Definition of nn:

nn is the count of the number of members in the consumer unit who have imputed data (whether due to invalid blanks, brackets, or both).

E. Topcoding Income Imputation.

The five income imputations and mean are all subject to the same topcoding rules as in previous releases of the public-use microdata. One critical value will be determined for all five imputations and the mean. If any of the imputations (INCOME1-INCOME5) for a particular member of the consumer unit fall above the upper critical value (for positive numbers) or below the lower critical value (for negative numbers), then that value will be topcoded and the associated flag will be coded as 'T'. Additionally, the mean value (INCOMEM) will be topcoded and the flag will also be given a value of 'T'.

For further information, see section IV of the Interview or Diary Documentation on "Topcoding and other nondisclosure requirements")

II. Applications.

A. Computing Means.

A1. Unweighted means.

As noted in the text, the mean income for a group of interest can be calculated by summing all data observations for the five imputations, and dividing by the total number of observations. Mathematically, the formula that applies is:

$$(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij})/(n \times m)$$

where X is the value of income, n is the number of rows, and m is the number of columns.

As an applied example, consider the following:

INTEARNX	INTEARNX1	INTEARNX2	INTEARNX3	INTEARNX4	INTEARNX5
100	100	100	100	100	100
D	50	250	300	20	80

In this example, the first consumer unit has reported a value for INTEARNX ($100), but the second consumer unit has only reported receipt of this income source. However, values (INTEARNX1 through INTEARNX5) are imputed for this consumer unit.

To find the mean value for the complete data set (i.e., the collected data and the imputed data), sum each observation (100 + 100 + … 100 + 50 + …. + 20 + 80) and divide the resulting total (1200) by the total number of observations (n*m=5*2=10) to get a mean value of 120.

However, the same value results when the mean of each row is calculated, and the mean of those means is then found. Using the same example, the data would now appear as follows:

INTEARNX1	INTEARNX2	INTEARNX3	INTEARNX4	INTEARNX5	INTEARNXM
100	100	100	100	100	100
50	250	300	20	80	140

Adding the two means (100+140) yields a total of 240. Dividing this by the number of means added (2) yields 120, the same value as obtained by finding the mean of all 10 observations.

A2. Weighted means.

In order to calculate the weighted mean without including variance calculations, the process using the complete data set is also straightforward. The weighted mean for the sixth column of data (INTEARNXM in this example) is calculated using the appropriate data weighting method described in the main text for which this documentation serves as supplement. The result is the weighted mean for this group. Specifically, suppose that the first consumer unit represents 5,000 similar units in the U.S. population, and the second consumer unit represents 7,500. In these circumstances, FINLTWT21 is 5,000 for the first consumer unit and 7,500 for the second unit. The weighted mean is: [(100*5,000) + (140*7,500)]/(5,000 + 7,500) or 124.

When variances are to be calculated, as described in the next section, it is recommended that the mean be found by calculating the mean of each of the five columns containing imputed data (that is, INTEARNX1 through INTEARNX5), and then averaging these means. Nevertheless, the mean will be the same, as demonstrated above for unweighted means. Following this procedure, the weighted means for each column are:

INTEARNX1	INTEARNX2	INTEARNX3	INTEARNX4	INTEARNX5
70	190	220	52	88

and the mean of these observations (70, 190, 220, 52, and 88) is 124.

B. Computing Variances.

When using multiply imputed data, the proper variance computation is straightforward, but involves more steps than the computation of variance from data sets in which no observations are initially missing. The reason is that in the latter case, all information is known. However, when data are imputed, there is additional uncertainty added to the complete data set by the very fact that the imputed data are estimates of values, rather than collected values. With multiple imputation, this imputation-related uncertainty is incorporated into the variance term, because more than one estimate of each missing value is posited. The proper variance is composed, then, of three elements: the "within imputation variance", which is the usual

variance computed for each column of the completed data set; the "between imputation variance", which accounts for variance across the columns of data; and an imputation adjustment factor described in Rubin (1987),[2] to account for the fact that a finite number of columns of data are created in the imputation process.

B1. Variances for Unweighted Means.

Consider the example shown in the section entitled, "Computing Means." In this case, two hypothetical consumer units reporting receipt of INTEARNX are shown, one of which reports a value ($100) while the other has values imputed. The data shown are:

INTEARNX	INTEARNX1	INTEARNX2	INTEARNX3	INTEARNX4	INTEARNX5
100	100	100	100	100	100
D	50	250	300	20	80

The first step is to compute the mean of each column of completed data (INTEARNX1 through INTEARNX5). Using notation consistent with Rubin (1987), this is:

$$\hat{Q}_{*_i} = \frac{\sum_{j=1}^{n} Q_{*ij}}{n} \qquad (1)$$

where Q_{*ij} is the nth observation of column i. In the current example, 1 <= i <= 5, and n = 2 for each column.

The next step is to calculate the average of the five complete data estimates $\overline{Q}_m$:

$$\overline{Q}_m = \frac{\sum_{i=1}^{m} \hat{Q}_{*i}}{m} \qquad (2)$$

where m is the number of columns containing multiply imputed data (i.e., m equals 5). Using the numbers above, $\overline{Q}_m$ is 120. (That is, it is the mean of the five column means, or the mean of 75, 175, 200, 60, and 90).

The third step is to calculate the variance of each column of data, using the standard variance formula:

$$U_{*i} = \mathrm{V}\left(\hat{\mathrm{Q}}_{*i}\right) = \sum_{j=1}^{\mathrm{n}} \left(Q_{*ij} - \hat{\mathrm{Q}}_{*i} \right)^2 \Big/ (n-1). \qquad (3)$$

The fourth step is to calculate the mean of these variances, or:

$$\overline{U}_m = \frac{\sum_{i=1}^{m} U_{*i}}{m} \qquad (4)$$

where $\overline{U}_m$ is the estimate of the *within* imputation variances. In the current example, the variances of the columns are found to be 1,250; 11,250; 20,000; 3,200; and 200. The mean of these values is 7,180.

[2] P. 84-91.

The fifth step is to calculate the variance *between* (or among) the five complete data mean estimates:

$$B_m = \sum_{i=1}^{m} \left(\hat{Q}_{*i} - \overline{Q}_m\right)^2 \Big/ \left(m-1\right) \qquad (5)$$

That is, B_m measures the variance of the means of each of the five columns. In the current example B_m is found to be 3,987.5, or the variance of 75, 175, 200, 60 and 90.

Now that the elements of the variance have been computed, the final step is to insert them into the formula for total variance (T_m):

$$T_m = \overline{U}_m + \left(1 + m^{-1}\right)B_m \qquad (6)$$

where $\left(1 + m^{-1}\right)$ is the imputation adjustment factor. Because there are 5 imputations in the completed data set, the factor is equal to 1.2. When all the elements are included in the equation, the variance of the unweighted mean (120) is computed to be 11,965 (that is, 7,180 plus 1.2 times 3,987.5).

B2. Variances for Weighted Means.

When calculating the variance for the weighted mean, the procedure is similar to the procedure for unweighted means. In this case, the weighted mean is used instead of the unweighted mean where appropriate. That is, continuing to rely on the example from the "Computing Means" section (in which the first consumer unit represented 5,000 similar units and the second represented 7,500), it can be shown that the five observations for each $\hat{Q}_{*i}$ are 70, 190, 220, 52, and 88, and that $\overline{Q}_m$ equals 124. Computing the variance, of each $\hat{Q}_{*i}$ is not easily shown, as it depends on the values of the 44 replicate weights. The method for computing these variances, though, is described in the main document to which this work is a supplement. That is,

$$U_{*i} = \mathrm{V}\left(\hat{\mathrm{Q}}_{*i}\right) = \frac{1}{44}\sum_{r=1}^{44}\left(\hat{\mathrm{Q}}_{ri} - \hat{\mathrm{Q}}_{*i}\right)^2$$

The formula for computing T_m is the same as described in the unweighted variance section, as is the computation of its elements ($\overline{U}_m$ and B_m).

C. Standard Error of the Mean (SE).

C1. Computation.

Once the total variance (T_m) is calculated, the standard error of the mean (SE) of the imputed data is calculated as usual—that is, $SE = \sqrt{T_m}$. Once obtained, the SE is used in the usual way in hypothesis testing. For example, the value can be used to compute a standard 95 percent confidence interval around the mean of the complete data set value of interest (that is, around $\overline{Q}$). However, the degrees of freedom associated with the t-value used in this computation are calculated according to a special formula described by Rubin (p. 77). See "Use in Hypothesis Testing," below, for details.

C2. Use in Hypothesis Testing.

As noted above, the SE can be used in standard hypothesis testing. For example, a standard confidence interval can be built around $\overline{Q}$ using SE in the conventional way. According to Rubin (p.77), the formula for the standard 100(1-α)% interval estimate of Q is:

$$\overline{Q} \pm t_{\upsilon}(\alpha/2)T_m^{1/2}$$

"where $t_v(\alpha/2)$ is the upper 100α/2 percentage point of the student *t* distribution on v degrees of freedom (e.g., if $v=\infty$ and $1-\alpha=.95$, $t_v(\alpha/2)=1.96$."

Note, though, that the value for degrees of freedom is calculated in a special way for imputed data. Again according to Rubin (p. 77),

$$v=(m-1)\left(1+r_m^{-1}\right)^2$$

where r_m is defined as the relative increase in variance due to nonresponse, and is computed according to the following formula:

$$r_m=\left(1+m^{-1}\right)B_m/\overline{U}_m\ .$$

In addition, Rubin (p. 77) provides the formula for computing an F-test in which $\overline{Q}$ is compared against a null value of interest, Q_0:

$$\text{Prob}\left\{F_{1,\upsilon}>\left(Q_0-\overline{Q}_m\right)^2/T_m\right\}$$

"where $F_{1,\upsilon}$ is an F random variable on one and v degrees of freedom."

C2a. Example: Computing a confidence interval

Using the example from the unweighted means section, recall that $\overline{Q}$ equals 120, and $T_m=11{,}965$. To compute the 95 percent confidence interval around this value, the formula described earlier in this section applies. That is,

$$\overline{Q} \pm t_{\upsilon}(\alpha/2)T_m^{1/2}$$

where

$$\overline{Q}=120;$$
$$T_m^{1/2}=SE=\sqrt{11{,}965}\approx 109.$$

To find $t_{\upsilon}(\alpha/2)$, the degrees of freedom must be calculated. As described earlier, B equals 3,9875.5, and $\overline{U}$ equals 7,180. Using this information, r_m is computed as follows:

$$r_m = \left(1 + m^{-1}\right) B_m / \overline{U}_m = \left[1 + \left(5^{-1}\right)\right] \times \left(3{,}975.5/7{,}180\right) = 0.664429 .$$

Thus, $v = (m-1)\left(1 + r_m^{-1}\right)^2 = (5-1)\left[1 + \left(0.664429^{-1}\right)\right]^2 \approx 25.10$. The t-value for the 95 percent confidence level with 25 degrees of freedom is approximately 2.06. Therefore, the confidence interval is computed as follows:

$$\overline{Q} \pm t_{\upsilon}(\alpha/2) T_m^{1/2} = 120 \pm (2.06 * 109))$$

The resulting confidence interval is:

$$-105 \le \overline{Q} \le 345 .$$

C2b. Example: Conducting an F-test

As described earlier, the F-test is used to compare $\overline{Q}$ to a null value. For example, suppose that the rest of the population reports average interest income of $150. To test whether or not the mean of the test sample ($120) is statistically significantly different from $150, the F-test is carried out as follows:

$$\text{Prob}\left\{F_{1,\upsilon} > \left(Q_0 - \overline{Q}_m\right)^2 / T_m\right\} \Rightarrow \text{Prob}\left\{F_{1,25} > (150-120)^2 / 11{,}965\right\} \Rightarrow \text{Prob}\left\{F_{1,25} > 0.075\right\}$$

At the 95 percent confidence level, $F_{1,25} = 4.24$. Because 4.24 is greater than 0.075, the null hypothesis is not rejected.

D. Distributional Analyses using Imputed Income

Currently, the Consumer Expenditure Survey publishes two standard tables that describe income class: range (e.g., less than $5,000) and quintile. Using these classifications, at least two different types of analysis can be performed: one where other characteristics are described as a function of, or related to, income; and one in which income distribution alone is of interest. An example of the first case is the current standard table publication. That is, these tables show how expenditures, age of reference person, and other characteristics differ across income classifications. An example of the second case is computation of the Lorenz curve or Gini coefficient for a particular group. (Examples of each follow.) The first method is used to produce the standard published tables, and is called the "publication method" throughout the remainder of this section. The second method is called the "distributional method."

D1. Publication Method.

In the standard published tables, income is used as a classifier variable, and means for expenditures, age of reference person, and other variables are described by income class (for example, less than $5,000 or first income quintile). With imputed income, the values in the "mean" column (i.e., the values for FINCBTXM) are used to classify consumer units by income group. This is because FINCBTXM represents the "best guess" of income for the consumer unit. As an example, suppose that the following observations are selected for study:[3]

CU	FINCBTX1	FINCBTX2	FINCBTX3	FINCBTX4	FINCBTX5	FINCBTXM

[3] These data are not actual imputed data. They are simulated by starting with $50,000 and adding or subtracting a random value between $0 and $49,999 to ensure all simulated values are between $1 and $99,999. Whether the random number is added or subtracted is also randomly determined.

1	51,580	22,701	53,967	87,617	298	43,233
2	89,164	96,337	62,853	74,799	45,814	73,793
3	38,841	83,616	72,586	75,456	30,077	60,115
4	20,568	19,116	54,186	19,190	4,896	23,591
5	5,114	10,352	44,733	39,086	36,163	27,090
6	41,488	64,692	626	94,851	77,271	55,786
7	58,957	535	35,711	22,920	17,212	27,067
8	54,711	16,527	85,930	54,136	18,579	45,977
9	92,395	90,650	54,030	98,502	61,983	79,512
10	98,228	25,890	54,191	34,835	97,515	62,132

To compute means and variances for the $20,000 to $29,999 income group, consumer units 4, 5, and 7 are used. The unweighted mean income for this group is: (23,591 + 27,090 + 27,067)/3 = $25,916. To compute the variance for this income group, the method described in the variance section is used. That is, the variance U1 is calculated using the values from FINCBTX1 for this group (20,568; 5,114; and 58,957). Using the formulas described in the variance calculation section (section IIB.), the standard error of the mean for this group is $21,413. To compute weighted means and standard errors, these same consumer units and their appropriate weights would be used in the way described in the variance computation section.

D1a. Quintiles.

Using these data, the mean and variance for each quintile can also be calculated. Because there are 10 observations shown here, each (unweighted) quintile is composed of two consumer units. To find the mean income for the first quintile, the data are sorted by FINCBTXM, and the first two consumer units (i.e., the first 20 percent in line) are selected for analysis. The resulting data set is:

CU	FINCBTX1	FINCBTX2	FINCBTX3	FINCBTX4	FINCBTX5	FINCBTXM
4	20,568	19,116	54,186	19,190	4,896	23,591
7	58,957	535	35,711	22,920	17,212	27,067

Mean income for this quintile is $25,329. The standard error of the mean for this quintile is $23,437.

D1b. Regression.

In regression analysis, it may be useful either to use income category as a binary variable, or to run separate regressions by income group. (For example, to calculate marginal propensity to consume food for the $20,000 to $29,999 group.) In these cases, the same classifications would be used as just described: that is, for consumer units 4, 5, and 7, the binary variable is equal to 1, and is equal to 0 for all other consumer units. If income is to be used as a continuous variable for the $20,000 to $29,999 group, then five regressions are run using FINCBTX1 through FINCBTX5 as described in the regression section.

D2. Distributional method.

At times, only mean income per group, and not variance, is needed. Two related applications of this case involve tools used in analysis of income distribution: The Lorenz curve and the Gini coefficient. (See section D2c. for details.)

To compute means by quintile in this case, each income variable (FINCBTX1 through FINCBTX5) is sorted by its associated INC_RANK value (that is, FINCBTX1 is sorted by INC_RANK1, etc.). Within each column, consumer units are divided into the appropriate quintiles, based on their INC_RANK group. Means are then calculated column by column for each quintile. The mean for each quintile derived from each column can be averaged as appropriate to derive the estimated mean for the quintile under study. For

example, using the data shown earlier, the unweighted mean for the first income quintile would be found as follows:

1. Sort each column in ascending order of income.

FINCBTX1	FINCBTX2	FINCBTX3	FINCBTX4	FINCBTX5
5,114	535	626	19,190	298
20,568	10,352	35,711	22,920	4,896
38,841	16,527	44,733	34,835	17,212
41,488	19,116	53,967	39,086	18,579
51,580	22,701	54,030	54,136	30,077
54,711	25,890	54,186	74,799	36,163
58,957	64,692	54,191	75,456	45,814
89,164	83,616	62,853	87,617	61,983
92,395	90,650	72,586	94,851	77,271
98,228	96,337	85,930	98,502	97,515

1. Select the first two rows of this table. These rows contain the data for the first 20 percent of the income observations within each column.

FINCBTX1	FINCBTX2	FINCBTX3	FINCBTX4	FINCBTX5
5,114	535	626	19,190	298
20,568	10,352	35,711	22,920	4,896

1. Sum the 10 values shown and divide by 10. The result ($12,021) is the unweighted mean for the first quintile.

A similar procedure is followed when deriving mean income for a particular income range. For example, to calculate unweighted mean income for the $70,000 and over group, observations from each column that fit this description are selected. For FINCBTX1 and FINCBTX2, the last three observations shown in the table in step 1 are selected. From FINCBTX3 and FINCBTX5, only the last two observations are selected. From FINCBTX4, the last 5 observations are selected. Averaging these values yields the mean ($87,661) for the $70,000 and over group.

In each case, weighted means are derived after applying the appropriate weights and following similar procedures.

D2a. Variances.

Note that no method for producing variances is described here. The reason is that each column of imputed income data can have a different number of observations when this method is used. As noted, the number of observations for mean income for the $70,000 and over group ranges from two (for FINCBTX3 and FINCBTX5) to 5 (for FINCBTX4). This difference in number of observations per column was not a factor in calculating the unweighted quintile values, but could be when weights are applied. Because the number of observations per column differs, the degrees of freedom for the calculation of variance as described by Rubin is no longer valid. In addition, note that the same consumer unit can appear in each of the five quintiles (because each of the five imputed values could fall in a different quintile) or income range. Therefore, computing average expenditure or age by range or quintile when using this method is not as straightforward as it is in the publication method. In the example shown in the publication method, the mean expenditure for the first quintile is the mean of the expenditure by consumer units 4 and 7. However, in this method, the mean expenditure for the first quintile is the mean expenditure by consumer units 4 and 5 for the first column of data; mean expenditure of units 5 and 7 for the second column of data; and so forth. Once each of these individual means are found, they would be added and divided by 5 to get the estimated mean expenditure by the first quintile under this method.

D2b. Regression.

For regression analysis, a similar problem occurs. When creating a binary variable indicating income is $70,000 or greater, for instance, the first and second regressions would have three observations of 1, and seven of 0. But the third and fifth regressions would have two observations of 1 and eight of 0. If actually running regressions by income class, again, each regression as a whole would have a different number of observations. The computation of the parameter estimates and their standard errors as described in the regression section would be invalid due to the degrees of freedom problem already described.

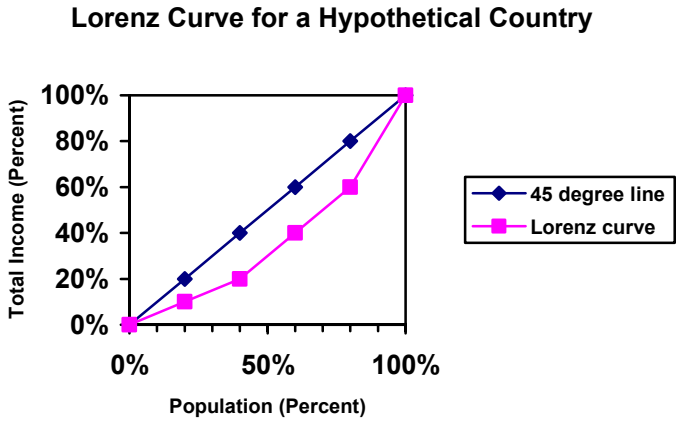
D2c. Examples: Lorenz curve and Gini coefficient.

The Lorenz curve depicts the percentage of total population income received by a particular percentage of the population. For example, the statement "Those in the lowest income quintile account for 20 percent of the population but receive 10 percent of total income in the country of interest" describes a point that would be depicted on a Lorenz curve. The Lorenz curve is usually compared to a 45 degree line, which indicates that income is equally distributed. (That is, at every point, X percent of the population controls X percent of income.) The Gini coefficient is the ratio of the area of the gap between the perfect equality (i.e., the 45 degree line) and the Lorenz curve to the total area under the perfect equality line. If there is perfect equality of income distribution (that is, all families or earners receive the same income), the gap between the two curves is zero, and therefore, the Gini coefficient is zero. If there is perfect inequality of distribution (one family has all the income in the country, to that the 100^{th} percentile controls 100 percent of income, but the 99^{th} percentile controls zero percent), the area of the gap equals the area under the perfect equality line, and the Gini coefficient equals 1.

The following table describes the data used to derive a Lorenz curve and Gini coefficient for a hypothetical country.

Population	Income
0%	0%
20%	10%
40%	20%
60%	40%
80%	60%
100%	100%

According to this table, 20 percent of the population of this country receives 10 percent of the income. These data can be depicted graphically as follows:



The area under the 45 degree line is 5,000 square percentage units (because the area of the triangle under the 45 degree line is {0.5*[100 percent*100 percent]}). The area between the 45 degree line and the Lorenz curve is 1,400 square units. The Gini coefficient is 1,400/5,000 = 0.28.

E. Computing Regression Results.

In order to use the multiply imputed income data in a regression framework and to calculate the mean and variance of the estimated coefficients, use repeated-imputation inference (RII). The proper estimation uses all five imputations for income by estimating the regression model once with each imputation. The procedure described applies to both weighted and unweighted regression analyses.

Note: This section uses examples specific to Ordinary Least Squares (OLS) regression. However, the process used to compute the OLS estimates from multiply imputed data sets generalizes to other types of regression, such as logistic regression.

A linear regression model is used for the formulas and for the empirical example. To begin, there is a dependent variable, y, and a vector of independent variables, x. For simplicity, assume a linear model is run for complete income reporters only on an intercept, before-tax income (FINCBTAX), and one other independent variable:

$$y = \alpha + \beta(\text{FINCBTAX}) + \gamma X + \varepsilon, \qquad (1)$$

in order to obtain estimates of the α, β, and γ. To obtain results using the imputed data for all consumer units, the regression model is estimated five times, once for each imputation:

$$\begin{aligned} y &= a_1 + b_1(\text{FINCBTX1}) + g_1 X, \\ y &= a_2 + b_2(\text{FINCBTX2}) + g_2 X, \\ y &= a_3 + b_3(\text{FINCBTX3}) + g_3 X, \qquad (2) \\ y &= a_4 + b_4(\text{FINCBTX4}) + g_4 X, \text{ and} \\ y &= a_5 + b_5(\text{FINCBTX5}) + g_5 X, \end{aligned}$$

To obtain the point estimates for each coefficient, calculate the mean of the five coefficients. For the slope coefficient on income, calculate:

$$\bar{b} = \frac{\sum_{i=1}^{m} b_i}{m} \qquad (3)$$

where *m* equals the number of imputations (five in this case). Similarly, to calculate the best point estimate for the intercept or slope coefficient on X, calculate the mean of the five estimates (a_1, a_2..., a_5 or g_1, g_2..., g_5).

To obtain the variance of the point estimate b-bar, the formula is identical to the formula given in the previous section. As a reminder the formula for the total variance (T_m) is:

$$T_m = \overline{U}_m + \left(1 + m^{-1}\right)B_m \qquad (4)$$

where T_m is the total variance of the coefficient, *U-bar* is the within imputation variance, and B_m is the between imputation variance.[4] The formulas for *U-bar* and B_m are:

$$\overline{U} = \frac{\sum_{i=1}^{m} U_i}{m}$$

where U_i is the variance of the estimated coefficient for imputation *i*. In other words, U-bar equals the mean of the five estimated variances. And,

$$B_m = \frac{\sum_{i=1}^{m} \left(b_i - \overline{b}\right)^2}{m-1}$$

Like above, B_m is referred to as the between imputation variance because it takes into account the uncertainty involving the point estimate. Consequently, B_m equals the variance of the point estimates.

Once B_m and *U-bar* are estimated, the variance of the *b-bar* can be calculated using (4), and the standard error of the *b-bar* is the square root of *T*.

E1. Other Statistics of Interest.

E1a. T-statistic.

To determine whether the point estimate is statistically different from zero, the simple t-statistic is calculated as the point estimate divided by the standard error.

E1b. F-statistic – single linear constraints.

To test whether the coefficient, *b-bar*, equals a constant, b_o, use an F-statistic:

$$F = \frac{\left(\overline{b} - b_0\right)^2}{T}$$

[4] The between imputation is weighted by the term in parentheses because we use a finite number of imputations. As m approaches infinity, the adjustment factor approaches one.

with one and v degrees of freedom, where v equals:

$$v = (m-1)\left(1+\frac{1}{r_m}\right)^2$$

and, r_m is the ratio of the between imputation and the within imputation variance, again weighted by the adjustment factor:

$$r_m = \frac{(1+1/m)B_m}{U}$$

E1c. Variance/Covariance matrix.

The variance/covariance matrix is calculated just like the variance. There is now a kxk matrix for B_m, where k is the number of independent variables plus the intercept (three in this example).

$$B_{\mathrm{m}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}(b_i-\bar{b})(b_j-\bar{b})}{m-1}$$

And, the kxk matrix for U is calculated the same way, where each element is the average of the five elements in each imputation's variance/covariance matrix.

E2. Numerical Example.

The following example derives from a regression of ZTOTAL on the variables described below.

	Coefficient	*Std. error*	*Variance*	*t-statistic*
First imputation				
Before-tax income	0.078	0.001	1.99e-06	55.06
Family size	808.171	62.574	3915.544	12.92
Intercept	3923.866	181.290	32866.209	21.64
Second imputation				
Before-tax income	0.078	0.001	2.04e-06	54.70
Family size	801.539	62.735	3935.699	12.78
Intercept	3923.418	181.702	33015.435	21.59
Third imputation				
Before-tax income	0.078	0.001	2.05e-06	54.31
Family size	803.341	62.878	3953.622	12.78
Intercept	3958.626	181.993	33121.488	21.75

cont.	Coefficient	Std. error	Variance	t-statistic
Fourth imputation				
Before-tax income	0.079	0.001	2.07e-06	54.96
Family size	797.967	62.650	3924.999	12.74
Intercept	3896.088	181.527	32951.979	21.46
Fifth imputation				
Before-tax income	0.078	0.001	2.08e-06	54.46
Family size	795.397	62.849	3949.982	12.66
Intercept	3919.280	181.992	33120.942	21.54
RII technique				
Before-tax income	0.078	0.002	2.28e-06	51.74
Family size	801.283	62.970	3965.185	12.72
Intercept	3924.256	183.345	33615.315	21.40

The components of the variance are also as follows:

	U	**B**
Before-tax income	2.04e-06	2.00e-07
Family size	3935.969	24.347
Intercept	33015.211	500.087

Test whether the coefficient on before-tax income equals 0.1. The hypothesis is:

H_0: b-bar = 0.1
H_1: b-bar ? 0.1

The F-statistic = 208.07, with 1 and 362 degrees of freedom. The critical value for this F(1, 362) is approximately 3.8, suggesting that the hypothesis that the coefficient on income equals 0.1 is rejected.

References.

Rubin, Donald B. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons, 1987.