Cor1014 Documentation and proposed usage for Imputation Data

K. Sicinski 11/2/10

1. Missing Data in Income and Asset Variables

The pattern of income and asset questions in the Wisconsin Longitudinal Study (WLS) grew more complex over time. This section focuses on the most recent 2004-2007 wave of interviews. Differences between waves are discussed in Section 3.

The sequence of questions about a given asset or type of income generally opened with a screener asking whether the participant owned that asset or had that type of income. If so, the amount was asked next. In case the amount was not given, an attempt to narrow the range of possible values by asking a series of "Would it amount to less than $XXX or more than $XXX" followed. The entry point into the series was chosen at random, though the top amount could never be asked first. The participant could terminate the bracketing process at any time by answering "Don't know" or refusing to provide an answer. Refusals in bracketing were counted and after the second occurred a conversion attempt was undertaken. If it was not successful, respondents unwilling to disclose any financial information were no longer asked for any amounts in the remainder of the section and coded as "Global refusal" in the data. For more details on the conversion protocol, see COR881.

The unfolding bracket design was pretty complex. The goal was to place bracket boundaries at the 10^{th}, 50^{th}, 80^{th} and 95^{th} percentile of the relevant amount distribution. When the distributions differed considerably by gender, separate values were chosen for male and female respondents. Initially the amount distributions were derived from the 1992 WLS and the Health and Retirement Study (HRS) data. Once half of the interviews were fielded, the distributions were reexamined using actual data. In some cases, this led to revisions of bracket boundaries. As a result, the number of brackets for a given amount variable can be quite large. For additional information on bracketing, see COR881.

There are a number of possible outcomes for the sequence of asset and income questions. An example of a variable summarizing them for spousal business income is provided below.

Response summary for gp204sp	Freq.	Percent	Cum.
GLOBAL REFUSAL	8	0.11	0.11
ITEM NOT ON PATH	3,154	43.41	43.52
DK/R SCREENER QUESTION	92	1.27	44.79
NO SUCH INCOME	3,643	50.14	94.93
AMOUNT GIVEN	201	2.77	97.70
PARTIAL INTERVIEW - QUESTION NOT ASKED	74	1.02	98.72
AMOUNT NOT ASCERTAINED	3	0.04	98.76
NO INFO IN BRACKETING	14	0.19	98.95
PARTIAL INFO IN BRACKETING	14	0.19	99.15
COMPLETE INFO IN BRACKETING	62	0.85	100.00

Total	7,265	100.00

The outcomes range from full information about the case to not even knowing whether the spouse had that type of income. Complete information was obtained when the item was not on path (not married or spouse retired,) no such income was received or the amount was given. Next in order of decreasing informational content are cases where the entire bracketing sequence was completed, followed by ones where some, but not all bracketing questions were answered. If there was a global refusal, the amount was not ascertained or no information was obtained in bracketing, only the fact that the income was received is known. In partial interviews and when the respondent refused or did not know the answer to the screener question, the data contain no information about the spouse's business income.

2. Imputation Process

Multiple imputations of missing wealth and income items was performed for the survey years 1975, 1992 and 2004 for graduates and 1994 and 2005 for siblings. The imputations were created using IVEware,[1] free SAS software which uses a sequential multivariate regression procedure.[2] IVEware automatically selects the appropriate regression model based on the properties of variable being imputed, using all other variables as potential covariates. The imputations are then drawn from the posterior predictive distribution of the chosen regression model. While the software imputes one variable at a time, the procedure can be repeated multiple times with previous imputations being overwritten in every cycle. This iterative process allows for interdependence to build among imputed values and better exploits the correlational relations among covariates.

IVEware imputes every variable in the data set, including covariates. For this reason, only an essential set of covariates was used. Family background controls included a factor-weighted SES score for parents, the respondent's perception of the family's economic status in 1957, religious background, population of the parents' place of residence, farm background, the number of siblings and an indicator for single-parent household. Respondent characteristics included gender, years of education, high school rank, IQ score, cognition score from 2004, health status, experience and its square, hours worked, tenure, the number of children, spousal education and indicators for marital status, full-time employment, farming, retirement status and self-employment.

IVEware allows the user to place restrictions on the imputed variables. This feature was used to introduce conditioning present in the interview instrument. For example, wages were imputed only if the participant was not retired, spousal variables were not imputed for single respondents. Most importantly, amounts were only imputed if the corresponding screener had an affirmative answer or was itself imputed with a "Yes". Thus, if the screener was missing, it is quite feasible for an imputed amount to exist in one set of imputations (screener imputed with "Yes") and not exist in another (screener imputed with "No"). The software also allows for the imputed values to be bounded, either be a constant or another variable. Lower and upper bounds for each amount variable were derived from the information collected in bracketing. For full information cases the bounds are often tight, when no information was collected the bounds were set at zero and infinity and thus imposed no restrictions on the imputed value.

After IVEware was run, some cleaning of the data was required. Valid missing values had to be reintroduced, as the software assigns numerical codes to inapplicable responses. All dollar amounts were

[1] Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2002). IVEware: Imputation and Variance Estimation Software. Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan. http://www.isr.umich.edu/src/smp/ive/

[2] Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, 27, 85-95.

transformed into logarithms before being imputed, this operation was reversed. Finally, constructed variables, such as equity or total personal income were created using the imputed data.

3. Differences Between Waves

Prior to 2004, the WLS did not use unfolding brackets. For those earlier waves imputations were only bound to be greater than zero. The second major difference come from the way income questions were asked: generally information on whether the respondent had a particular type of income was not collected. Instead, the amount questions were asked directly with zeros indicating no such income. In such cases artificial screeners were created in order to prevent every missing value from being imputed with a nonzero amount. The artificial screeners were missing whenever the respondent did not provide an amount; a value was imputed conditional on the screener being imputed with a "Yes", otherwise the imputation is zero. While personal characteristics were updated with wave-appropriate values, the set of covariates was largely unchanged between waves. Fewer characteristics were used for siblings, as the survey contains less information about them. Occasionally IVEware would produce segmentation faults. In such circumstances minor changes in the setup parameters were made to eliminate the problem.

4. Finding and Using the WLS Imputations

To facilitate the assessment of variability due to imputation, five sets of imputed values are available in the WLS. An imputed variable is identified by the suffix "i**n**", where **n** is the set identifier. For example, the first imputation of spousal business income (gp204sp) for the 2004 wave would be gp204spi1 and the fourth would be gp204spi4.

There are several arguments for using the WLS imputed data. First, it accounts for observed differences between complete and missing cases. Second, it prevents loss of information that occurs if cases with a missing value are discarded. Third, it provides consistent treatment of the nonresponse problem for all users. While the above benefits apply to any of the five imputed data sets, the full benefit of imputations is realized when multiple imputation techniques are used. The disadvantage of using a single imputation in statistical analysis stems from the fact that standard software will not correct for the increased sample size that is due to imputation. This will result in downward biased standard errors, overly tight confidence intervals and too frequent rejection of the null hypothesis in significance tests. These problems are overcome when the estimation technique accounts for the fact that an imputation is only a plausible value and not the actual response, as is the case in multiple imputation.[3] Newer statistical packages often contain tools for analyzing multiply-imputed data. Should these not be available, the necessary adjustments can still be done by hand. This involves analyzing the five sets separately using standard methods and then applying a formula that corrects the results for uncertainty due to imputation. For more detail, see Rubin (1987).

[3] Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.