INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
PANEL STUDY OF INCOME DYNAMICS
UNIVERSITY OF MICHIGAN

Technical Series Paper #20-01

Multiple Imputation with Massive Data: an Application to the Panel Study of Income Dynamics

July 19, 2020

Yajuan Si[1*]
Steve Heeringa[1]
David Johnson[1]
Roderick Little[2]
Wenshuo Liu[3]
Fabian Pfeffer[1]
Trivellore Raghunathan[2]

1. Survey Research Center, University of Michigan, Ann Arbor, MI, USA; 2. Department of Biostatistics and Survey Research Center, University of Michigan, Ann Arbor, MI, USA; 3. Interactions LLC.

*Corresponding email: yajuan@umich.edu

Abstract

Multiple imputation (MI) is a popular and well-established method for handling missing data in multivariate data sets, but its practicality for use in massive and complex data sets has been questioned. One such data set is the Panel Study of Income Dynamics (PSID), a longstanding and extensive survey of household income and wealth in the United States. Missing data for this survey are currently handled using traditional hot deck methods. We use a sequential regression/chained-equation approach, using the software IVEware, to multiply impute cross-sectional wealth data in the 2013 PSID, and compare analyses of the resulting imputed data with results from the current hot deck approach. Practical difficulties and our approaches to overcoming them, are described in this setting. We evaluate the imputation quality and validity with internal diagnostics and external benchmarking data. Though MI produces some improvements over the existing hot deck approach, gains are limited due to a relatively small fraction of missing information in this application. We demonstrate the practical implementation and expect greater gains when the fraction of missing information is large.

Keywords: multiple imputation; massive data; validity; diagnostics.

1. Introduction

Multiple imputation (MI) is a useful tool for dealing with missing data, given its attractive theoretical properties, its ability to handle any pattern of missing data, and the numerous computation platforms that are available in practice. Since the initial development by Rubin (1987), MI has been successfully applied in a variety of fields for missing data and more broadly to handle related problems such as measurement error, confidentiality protection, and finite population inference (Reiter and Raghunathan, 2007; Van Buuren, 2012; Carpenter and Kenward, 2013).

The keys to success for MI are models that use available information on observed values of variables that are predictive of the missing values. Grounded in Bayesian methodology, MI draws model parameters from their posterior distributions and then imputes draws from a posterior predictive distribution or using predictive mean matching. The result is M completed datasets, each with different draws or imputations of the missing values. Variance estimation combines the within-imputation and between-imputation variance across these M datasets, using simple MI combining rules (Rubin 1987). Although its etiology is Bayesian, MI has been shown to yield efficient estimates and inferential validity from the frequentist perspective (Rubin, 1996; Meng, 2002).

Various MI software packages have been developed based on joint multivariate normal distributions, e.g., PROC MI (SAS Institute Inc., 2017), Amelia (King et al., 2001) and norm (Schafer, 1997), or a sequence of fully conditional distributions, e.g., IVEware (Raghunathan et al., 2001), MICE (Van Buuren and Oudshoorn, 1999), and MI (Gelman et al., 2015). MI for multilevel models is available in the R packages REALCOM-IMPUTE (Carpenter and Kenward, 2013) and PAN (Schafer, 2016) . Flexible nonparametric Bayesian mixture models are also applied to jointly impute a large number of incomplete categorical variables and a mixed group of categorical and continuous variables, e.g., NPBayesImputeCat (Si and Reiter, 2013) and MixedDataImpute (Murray and Reiter, 2016). Other prediction algorithms applied to chained equation MI include classification and regression trees (Burgette and Reiter, 2010), Bayesian additive regression trees (Xu et al., 2016) and random forests (Stekhoven and Bühlmann, 2012).

The big data era has led to increased availability of massive datasets with large sample sizes, and many variables with complex dependency structures. Loh et al. (2019) argued that MI under parametric models cannot be successfully implemented in these settings. Others (e.g. Little 2020) have contested this assertion. Though Stuart et al. (2009) and He et al. (2010) successfully applied chained-equation approaches to large data sets, our application is considerably more complex than that considered by Loh et al. (2019), who restrict attention to imputation of a single variable (amount of interest and dividend income) in the U.S. Consumer Expenditure Survey. This paper aims to illustrate some challenges in implementing MI on a large dataset by imputing an extensive set of variables in the Panel Study of Income Dynamics (PSID).

2. Motivating application

The PSID began in 1968 with a sample of over 18,000 individuals living in 5,000 U. S. families and it has followed them for the past five decades (PSID 2020). The PSID sample is dynamic and grows as children and grandchildren from these families form their own households and are recruited into the PSID sample and longitudinal data collection. One of the key study topics interested by researchers is the collection of wealth information for the households of these individuals and their descendants. The survey instrument contains questions on a range of separate wealth components, such as home values, mortgages, different types of financial assets, real assets, and debts; together, these components form a measure of net worth. Item non-response for most wealth components is quite low in the PSID (< 5%) partly due to the study's use of unfolding brackets to minimize nonresponse. However, about 20%–25% of families do not report a continuous value (e.g. either report a bracketed value or no value) for at least one of the components needed to compute net worth. Consequently, since the first wealth module in 1984, the PSID has imputed missing values for users and summed the imputed components to create aggregate net worth measures. We describe a project to multiply impute missing data in PSID's 18 wealth components, along with the missing values of predictors of these components. We considered a total of 409 variables with varying shares of missingness.

The current approach is a univariate hot deck imputation, where each case of item non-response (and bracketed response) is flagged and assigned a value that is randomly drawn from the set of reported values (or within the same bracket) with selection probabilities equal to the distribution of observed continuous values (or within the respective bracket) (for details see Pfeffer and Griffin (2017)). This approach has three fundamental limitations: the current hot deck approach does not condition on covariate information, it treats each source of wealth independently, and it does not allow the user to incorporate imputation uncertainty into estimates of standard errors. Our goal was to address these issues by applying MI, including as predictors various socioeconomic and demographic characteristics, the values of income and other asset components, and known predictors of wealth fluctuations such as changes in household composition, employment and retirement status, health conditions, and residence (Pfeffer and Griffin 2017).

We focus on cross-sectional PSID data for 2013 for 9,063 families using 409 variables that we selected to captures central socio-demographic characteristics of households, household reference persons and partners, such as employment, wages, family income, consumption, race, education, and wealth.

In datasets with structural zeros, where certain variables are “not applicable” given values of other variables, it is important to code variables in a way that distinguishes between “not applicable” and “missing.” For instance, in the case of wealth variables, PSID includes filter questions that ask whether a household holds a certain asset or not, such as whether the household owns their home. To those indicating home ownership, a follow-up question is fielded to ascertain the value of the home and remaining mortgages. For those indicating no home ownership, the follow-up variable is recorded as “non-applicable”. Both the house ownership indicator and the house value variable may be missing.

Table 1 About Here

In the 2013 PSID wave, 213 out of the 409 variables we selected are incomplete, and the “apparent” missingness proportions – the share of cases without a valid value – vary between 0.01% and 99.72%, with a median value of 47.22%, and 40 variables have more than 80% missing values. However, these apparent missingness summary is misleading as it fails to distinguish meaningful missingess arising from “non-applicable” questions. Table 1, therefore, shows the apparent and true missingness proportions for all 18 wealth components. The latter values are much smaller than the former, indicating that most of the missing values arise from non-applicable cases. Figure 1 depicts the missingness patterns of the components for the 1,457 families with incomplete wealth information, showing neither a monotone or nested structure.

3. MI Methodology

The regression models that underlie MI software tools need to be carefully developed and specified by the user, based on exploratory data analysis, model building and model checking through residual diagnostics. With 213 incomplete variables of mixed types and with various structural restrictions, deriving their joint distribution or developing a coherent joint imputation model is not feasible. In contrast, chained equation or sequential regression imputation approaches are both flexible and computationally attractive. This latter approach models the regression of each variable with missing values on all the other variables in the data set. It cycles iteratively through the dataset, imputing the missing values of each variable in turn, with missing values of predictors replaced by their most recent imputations. The cycles are similar to those of a Bayesian Gibbs' sampler, but the method is only approximately Bayesian because the sequence of conditional distributions may not correspond to a coherent joint distribution.

We now describe issues we encountered in creating MI datasets for PSID based on sequential regression imputation models using IVEware (Raghunathan, 2020).

3.1 Data preparation

Sample inclusion. We include all families who responded in the 2013 PSID. We draw on family-level and individual-level information as potential predictors of wealth components. The family-level predictors include income, consumption, and other financial measures, and the individual-level predictors include socio-demographics, including employment, wage and income information for household reference persons and partners (for a full list, see Appendix 1). As part of the chained equation, all components of wealth also become predictors of other components of wealth.

Recoding missing values. We impute missing values for all wealth measures and predictors. As mentioned above, different values (e.g., 999999, 0 or NA) indicating missing data are recoded as the same missing flags. PSID also provides flags for some variables to indicate changes to the originally reported value that resulted from editing or other processing steps. Cases that are edited during data cleaning are not considered missing. If the flag code shows the value is imputed by other methods, we recode it as missing. The "not applicable" values are treated as such and passed over in the imputation process via specified restrictions.

Outlier Detection. This usually has been done in the data cleaning or editing step. We plotted the frequency distributions of observed values to detect extreme values that would introduce skewness to the distributions of the individual variables.

Transformations. Most of the wealth variables, especially asset/debt amounts, are severely right-skewed, with a few very large values. After visually checking the frequency histograms of the observed values and calculating skewness parameters, we chose the cube root transformation for the wealth, income, and wage variables. This transformation substantially symmetrizes the shape of the distribution. Unlike the logarithm, the cube root transformation can be applied to negative and zero values, which occur for some variables in our study. As an illustration, Figure 2 shows the frequency histograms for the original home values (restricted to the sample families that report owning a house/apartment), after cube root transformation and logarithm transformation. The cubic root transformed values are approximately normal with a skewness of 0.7.

Figure 2 About Here

3.2. Developing the imputation models: Imputations are draws from the predictive posterior distribution of each missing variable, based on a regression model for each incomplete variable, preferably with all the predictors. Given a large number of predictors, we use forward variable selection to identify a subset of the predictors tailored to the variable type: specifically, linear for continuous variables, logistic or multinomial logistic for categorical variables, and Poisson for count variables. The wealth, income, and consumption variables are semi-continuous, a mix of a binary variable indicating presence or absence, and a continuous value if the variable is present. A two-stage model was used to impute missing values of these variables. First, a logistic regression model was used to impute presence or absence of an asset. Conditional on imputing a non-zero status, a normal linear regression model for the cube-root transformed outcome value was then used to impute non-zero values. For example, we first impute the indicator of whether the family has any real estate, and then impute the real estate value if owned. The indicators of non-zero status and amounts if non-zero then become potential predictors in imputation models for other variables.

In fitting these regressions in IVEware, multicollinearity issues were the main source of run errors. One reason was that as PSID has evolved with slightly different recoded versions of the same variables over time as well as versions of aggregated variables created based on user interests. Naturally the aggregated variables are collinear with their components. We applied principle component analyses to identify such variables and remove them from the set of predictors for the imputation models.

Another cause of multicollinearity were categorical variables with many nominal levels, such as the state of residence with 51 values. This variable is included as a predictor in the regression model with 51 dummy variables, leading to potential collinearity, especially from categories with small sample sizes, in this case, small states. We use forward selection to select the dummy variables along with other predictors to avoid issues with collinearity. Case identifiers, flag variables, and boundaries of intervals are not used as predictors during imputation but are used in defining whether the imputation should be done or not and for the boundaries for the imputed values.

All regression models were fitted only to the set of applicable cases; for example, the regression model for the imputation of house values was restricted to households that own their homes. Because homeownership can itself be imputed, the set of applicable cases changes in each IVEware iteration. These restrictions can be nested and must be explicitly specified so that the higher-level restricting variables are not used as predictors in the regression model. Restrictions also arise from nested skip patterns in the questionnaire. For example, a question about a loan for a second house is asked only when the respondent indicates first having a house and then a second house. Given specified restrictions, some variables are constrained to be positive, for example, the value of an owned house. However, other variables could still take on 0 values with restrictions, requiring a semi-continuous variable declaration. For example, a family could own real estate (restriction) but with or without a mortgage, where the mortgage value is a semi-continuous variable.

Some missing values come with logical or consistency bounds that must be accounted for during imputation. For example, the annual property tax or insurance premium amount must be non-negative. For some survey variables such as wealth and income, some respondents do not provide an exact value and instead answer follow-up questions asking for brackets or range for survey variable, which then define bounds within which the imputed values must lie. The bounds are incorporated by drawing imputations from a predictive distribution restricted to lie within the bounds, which is an option in IVEware.

Eliminating covariates that are not predictive of the outcome variable with missing values avoided problems with multicollinearity and helped to improve the convergence of the IVEware iterations. IVEware has the ability to use the marginal increase of the goodness of fit statistic R^2 when including each variable to select the variable. We set a minimum R^2 increase of 0.005 and a maximum number of predictors of 10.

The forward variable selection method to determine the models is admittedly rather ad hoc, but perhaps justified given the size and complexity of the problem, the need to avoid collinearities, and the goal of prediction of the missing values rather interpretation as in a substantive model. In principle, more sophisticated methods such as ridge regression or lasso could be implemented (see e.g. Deng et al., 2016), although doing so would be challenging in this particular missing data setting. Often additive models are the default option in regression models, but interactions that are potentially predictive can and should be explicitly included as covariates.

3.3. Model Diagnostics

It is important to check that imputations are plausible as unanticipated problems in setting up the regressions can lead to poor imputations. Imputed and observed values can be compared using graphical and numeric diagnostic tools (Stuart et al., 2009). The marginal distributions of observed and imputed values are expected to be similar under Missing Completely at Random (MCR) but may markedly differ under missing at random (MAR) conditions. Nevertheless, comparisons across the imputed data sets will be useful as a first phase of evaluation. Table 2 examines the summary information of total wealth and compares observed and completed datasets. The summary statistics of the completed values are generally larger than those of the observed, but not dramatically so.

Table 2 About Here

Bondarenko and Raghunathan (2016) developed graphical and numeric diagnostic tools for MI to compare the distributions of imputed and observed values conditional on the response propensity score. However, such tools need extension to work for imputation with restrictions and semi-continuous variables. A useful feature of MI is that the fraction of missing information (FMI), which estimates the relative increase in variance due to missing data (Rubin, 1987; Raghunathan, 2016), is readily computed as a simple function of the between-imputation and within-imputation variance. Note that no comparable measure is available from a single imputation method. A large value of FMI indicates substantial efficiency loss due to nonresponse. Table 3 presents the FMI values of filter and amount variables for the 18 PSID wealth components. Most of them are very small, primarily because of the low underlying missingness proportions. The last three columns in Table 3 compare the percentages between the observed and imputed values of positive indicators of whether the PSID household owns a wealth component and lists the number of cases with missing indicators. The comparison does not raise red flags for most components, except for the indicator A23 (Do you have a mortgage or loan on this property?) for A24 1 (Remaining principal of the first mortgage). In the imputation model for A23, the selected predictors include the house type (such as a one-family house, a two-family house, an apartment, a mobile home, or others), total family income, wage of the household head, indicator of credit card debts and amount of the saving account. We do not have substantive concerns about the plausibility of the imputed values, given that only 23 values are missing.

Table 3 About Here

3.4. Creating Imputations from IVEware multiple chains: The number of cycles in the initial burn-in period, the number of iterations between creating one set of imputations, and the number *M* of imputations to be performed, together determine the total number of cycles of the Gibbs' type algorithm. Comparisons of results for different choices of the number of cycles suggested that about 10 cycles were sufficient for most imputations.

4. Comparisons of Imputations from MI and from the Current Hot Deck Procedure

To assess the validity of our imputations, we compare the imputed values produced by the existing hot deck approach and those produced by our MI approach in three ways: We assess bivariate associations between wealth components as well as between net worth and other economic correlates. We then assess the performance of our newly imputed net worth variable as a prediction outcome. Finally, we compare the distribution of our newly imputed net worth variable to external, gold-standard data.

4.1. Bivariate Associations

The current hot deck imputation method performs univariate imputations, ignoring relationships with other variables, although other hot deck imputation approach can be done within adjustment cells. Chained equation MI, on the other hand, preserves the dependency structure by including predictive covariates in the imputation model. We first examine this MI improvement by considering the bivariate associations between wealth components. The existing imputation procedure generates the home equity value differently from MI (defined as A20 minus A24_1 minus A24_2) without reporting the three related components. This is another advantage of MI—the ability to handle hierarchical restrictions that could be also missing. Hence, we use the final home equity values, rather than the three components, and compare the resulting 16 wealth components between the two imputation methods. Figure 3 is a scatterplot of the pairwise Pearson correlation coefficients between the 16 wealth components, based on one MI dataset and the dataset imputed via hot deck. MI generates larger values than the hot deck method, suggesting that it is preserving associations better.

Next, we evaluate the association of the household net worth with household income, age, and education. Figure 4 marks the data grouped by income quartiles and compares the imputed wealth values from the hot deck and one MI dataset. The relationship between income and wealth is higher after MI than after hot deck imputation. As this relationship is expected to be positive (Kilewald et al, 2017), this comparison favors the MI method.

Table 4 examines the 10 imputed datasets and takes the average of the Pearson correlation coefficients between total net worth and total household income, age and education, respectively. We use the cube root transformed values for both wealth and income to dampen the effects of large outlying values. MI preserves the dependency structure from the observed data, while the univariate hot deck imputation attenuates the correlation estimates. We also compare the regression coefficient estimates of these sociodemographic predictors in the univariate regression models with net worth as the outcome and find that the relationships in MI are larger and closer to the observed structure than those after the hot deck imputation.

Table 4 does not report the corresponding variance estimates on either the observed or imputed dataset. Because the missingness proportion is small, inferences based on the completed datasets are similar between the two imputation methods. We use Table 4 to exemplify the different imputation methods' properties for preserving relationships among the variables that are being imputed. We checked the relationship between total household income with the 18 wealth components, respectively, and MI yields correlation coefficient estimates closer to those based on the observed values than the current hot deck method.

Table 4 About Here

4.2. Multivariate regression

We fit a multivariate regression model with household net worth as the outcome, and covariates that include total household income, education, race/ethnicity, age and marital status. Table 5 shows the coefficient estimates based on the observed values as complete case analysis, and the completed datasets after MI and hot deck imputation. We applied combining rules to the ten MI datasets to propagate the missing data uncertainty. The results under the three methods are similar, likely because the missingness proportions are small. We calculated the ratio of the MI variance estimates to those from hot deck (HD) imputation, the overall variance and the average within-imputation variances. Even with the additional variance component accounting for missing data uncertainty, the overall variances under MI are slightly larger than those under hot deck imputation but with smaller within-imputation variances, suggesting efficiency gains within imputation.

Table 5 About Here

4.3 Comparison with external data

In addition to the internal checks of imputation plausibility just presented, we compare our imputation results to external data. We contrast our estimates with another study, the Survey of Consumer Finances (Federal Reserve Board, 2020). The Survey of Consumer Finances (SCF) is a triennial cross-sectional survey of U.S. families and collects information on families' balance sheets, pensions, income, and demographic characteristics. The SCF oversamples wealthy households, so we apply survey weights to the imputed datasets and generate population-representative estimates. Table 6 compares the mean values of key wealth components based on the 2013 PSID MI and hot deck imputed datasets after weighting and the weighted 2013 SCF data. Because the PSID oversamples low-income families, the weighted PSID wealth estimates are higher than the unweighted. The values from the imputation methods are similar but both are lower than the SCF estimates as the SCF oversamples high-income families (Pfeffer et al. 2016).

Table 6 and 7 About Here

Since the wealth components have right-skewed distributions, we compare the percentiles for key distribution points: the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles, shown in Table 7. We have omitted the calculated percentiles of 0 values across all methods. Generally, the PSID estimates are lower than the SCF estimates, the latter of which have to adjust for the oversampling of wealthy families. MI generates lower values for checking/saving balances and stocks but larger values for IRS/private annuities and other debts than the hot deck imputation. The weighted estimates of business assets and other debts after MI are larger but the weighted estimates for other assets are lower than those produced by hot deck imputation.

5. Discussion

Chained-equation imputation methods are flexible, and can handle general missing data patterns of missing data. Imputations can be tailored to multiple variable types. We have discussed some of the methodological and practical issues encountered when we applied chained-equation MI in a large-scale, complex, particular application, the imputation of wealth data in the Panel Study of Income Dynamics.

Issues addressed include (a) distinguishing missing and nonapplicable values, and tailoring the underlying regression models to applicable cases, (b) transformations of amount variables to reduce skewness, (c) simultaneous imputation of presence and amount of semi-continuous variables, (d) minimizing collinearity in the regressions, and (e) model checking. The imputation models are tailored to variable types and the imputations account for restrictions and boundaries. Forward selection was applied to restrict the predictors to variables associated with the variable being imputed, speeding the convergence of the algorithm. Multiple sequences of conditional imputation models yield $M > 1$ completed datasets that can be used for standard analyses and MI inferences that propagate imputation uncertainty.

Little (2020) summarizes the important factors in this imputation setting, such as the sample size, fraction of missing information, missingness mechanism, form and strength of the true relationship between the variable with missing data and predictors, form and strength of the true relationship between missingness and its predictors, degree of association between the propensity to respond and the variable with missing values, and degree of misspecification of the true models for missingness and the survey design variables. We have described how we addressed each of these concerns in our imputations of PSID wealth data.

In this application, we compared MI with the existing single hot deck imputation method. To evaluate imputation plausibility, we examined descriptive summary statistics, bivariate association studies, multivariate regression models, and compared estimates with external estimates from the Survey of Consumer Finances. We demonstrated the capability of MI to preserve the data dependency structure and improve estimation efficiency.

This comprehensive investigation invites several extensions. First, we omitted the design information in the imputation process, and survey weights may be informative for the imputation process. However, the method to adjust for survey weights and design variables needs further investigation and software development efforts, such as models with random effects (Reiter et al., 2006) or penalized spline functions (Zheng and Little, 2003). Second, we focus on the 2013 cross-sectional PSID data. PSID's panel structure with repeated measures of the same family across time may further improve the imputation model with highly predictive variables, such as measures from previous waves of the longitudinal study.

References

Bondarenko, Irina, and Trivellore Raghunathan (2016). "Graphical and Numerical Diagnostic Tools to Assess Suitability of Multiple Imputations and Imputation Models." Statistics in Medicine 35: 3007–3020.

Burgette, L. F. and J. P. Reiter (2010). Multiple imputation via sequential regression trees. American Journal of Epidemiology 172, 1070–1076.

Carpenter, J. and M. G. Kenward (2013). Multiple imputation and its application. John Wiley & Sons.

Deng, Y., C. Chang, M. Ido, and Q. Long (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific Reports, 6, Article number: 21689.

Federal Reserve Board (2020). Survey of consumer finances. https://www.federalreserve.gov/econres/scfindex.htm.

Gelman, A., J. Hill, Y.-S. Su, M. Yajima, M. Pittau, B. Goodrich, Y. Si, and J. Kropko (2015). Package "mi". R CRAN.

He, Y., A. M. Zaslavsky, and M. B. Landrum (2010). "Multiple Imputation in a Large-Scale Complex Survey: A Guide." Statistical Methods in Medical Research 19: 653–670.

King, G., J. Honaker, A. Joseph, and K. Scheve (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Review 95, 49–69.

Little, R. (2020). On algorithmic and modeling approaches to imputation in large data sets. Statistica Sinica, forthcoming.

Meng, X. L. (2002). A congenial overview and investigation of multiple imputation inferences under uncongeniality. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. Little (Eds.), Survey Nonresponse, pp. 343–356. John Wiley & Sons.

Murray, J. S. and J. P. Reiter (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. Journal of the American Statistical Association 111 (516), 1466–1479.

Pfeffer, F. T. and J. Griffin (2017). Determinants of wealth fluctuation: Changes in hard- to-measure economic variables in a panel study. Methods, data, analyses 11 (1), 87–108.

PSID (2020). The Panel Study of Income Dynamics. https://psidonline.isr.umich.edu.

Raghunathan, T. (2020). IVEware: Imputation and variance estimation software, version 0.3. https://www.src.isr.umich.edu/software/.

Raghunathan, T. E. (2016). Missing Data Analysis in Practice. Boca Raton, FL: Boca Raton: CRC Press.

Raghunathan, T. E., J. M. Lepkowski, J. van Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27, 85–96.

Reiter, J. P. and T. E. Raghunathan (2007). The multiple adaptations of multiple imputation. Journal of the American Statistical Association 102, 1462–1471.

Reiter, J. P., T. E. Raghunathan, and S. K. Kinney (2006). The importance of modeling the survey design in multiple imputation for missing data. Survey Methodology 32, 143–150.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association 91, 473–489.

SAS Institute Inc. (2017). SAS/STAT 9.4 User's Guide.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schafer, J. L. (2016). Multiple imputation for multivariate panel or clustered data. https://cran.r-project.org/web/packages/pan/.

Si, Y. and J. P. Reiter (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. Journal of Educational and Behavioral Statistics 38 (5), 499–521.

Stekhoven, D. J. and P. Bühlmann (2012). Missforest-nonparametric missing value imputation for mixed-type data. Bioinformatics 28 (1), 112–118.

Stuart, E. A., M. Azur, C. Frangakis, and P. Leaf (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. American Journal of Epidemiology 169 (9), 1133–1139.

Van Buuren, S. (2012). Flexible Imputation of Missing Data. CRC Press.

Van Buuren, S. and C. Oudshoorn (1999). Flexible multivariate imputation by MICE. Technical report, Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

Xu, D., M. Daniels, and A. G. Winterstein (2016). Sequential BART for imputation of missing covariates. Biostatistics 17 (3), 589–602.

Zheng, H. and R. J. Little (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. Journal of Official Statistics 19 (2), 99–107.

Table 1. Apparent missingness proportions (Apparent miss) and true missingness proportions (True miss) for the 18 wealth components in the 2013 PSID study.

	Label	Apparent miss (%)	True miss (%)
W28	AMOUNT OF ALL ACCOUNTS	37.77	5.10
W39A	AMOUNT OF CREDIT/STORE CARD DEBT	66.60	1.31
W39B4	AMOUNT OF LOANS FROM RELATIVES	98.68	0.39
W39B3	AMOUNT OF LEGAL BILLS	99.39	0.40
W39B2	AMOUNT OF MEDICAL BILLS	89.83	0.81
W2A	WORTH OF OTR REAL ESTATE	89.48	0.81
W2B	AMOUNT OF OWED ON OTR REAL ESTATE	89.25	0.57
W39B7	AMOUNT OF OTHER DEBT	98.82	0.40
W39B1	AMOUNT OF STUDENT LOANS	75.47	1.20
W22	VALUE OF IRA/ANNUITY	81.82	2.52
W11A	WORTH OF FARM OR BUSINESS	92.79	1.39
W11B	AMOUNT OF OWED ON FARM OR BUSINESS	92.08	0.67
W34	PROFIT IF SOLD BONDS/INSURANCE	91.11	2.48
W16	PROFIT IF SOLD NON-IRA STOCK	89.44	2.01
W6	PROFIT IF SOLD VEHICLES	4.63	4.63
A20	HOUSE VALUE	50.40	1.48
A24 1	REM PRINCIPAL MORTGAGE 1	66.66	2.28
A24 2	REM PRINCIPAL MORTGAGE 2	95.90	0.54

Table 2: Summary statistics comparison between observed and completed 2013 PSID family wealth values based on one randomly selected multiple imputation dataset: sample size n, minimum/maximum, mean, standard deviation Std and quantiles. The relative difference (Rel.diff) is defined as (Com-Obs)/Obs. The wealth values are presented in $1000s and rounded to the nearest 1000.

	Obs	Com	Rel.diff
n	7606	9063	0.19
Min	-995	-995	0
Max	33740	33740	0
Mean	180	200	0.11
Std	793	837	0.06
25th	0	0	
50th	14	20	0.39
75th	114	135	0.18
90th	423	487	0.15
95th	846	912	0.08

Table 3: Fraction of missing information for wealth components and the corresponding indicators (Ind). YesProp_obs and YesProp_imp represent the proportions of positive indicators in the observed and imputed data, respectively. # missInd is the number of cases with missing indicators.

Component	FMI	Ind	Ind-FMI	YesProp_obs	YesProp_imp	#missInd
W28	0.002	W27	0.010	0.671	0.615	65
W39A	0.006	W38A	0.003	0.344	0.318	44
W39B4	0.000	W38BRELS	0.004	0.014	0.061	33
W39B3	0.011	W38BLEGL	0.003	0.006	0.000	33
W39B2	0.008	W38BMED	0.005	0.107	0.182	33
W2A	0.007	W1	0.002	0.111	0.190	21
W2B	0.012	W1	0.002	0.111	0.190	21
W39B7	0.002	W38BOTR	0.002	0.012	0.030	33
W39B1	0.014	W38BSTU	0.004	0.255	0.176	34
W22	0.010	W21	0.002	0.203	0.200	50
W11A	0.031	W10	0.001	0.084	0.059	17
W11B	0.024	W10	0.001	0.084	0.059	17
W34	0.013	W33	0.008	0.106	0.145	76
W16	0.012	W15	0.003	0.120	0.161	56
W6	0.013					0
A20	0.002	A19	0.000			0
A24_1	0.023	A23	0.002	0.355	0.739	23
A24_2	0.031	A28_1	0.011	0.043	0.143	35

Table 4: Correlation between total wealth (cubic root transformed) and sociodemographics.

	Obs	MI-imp	Hot deck-imp
Total family income	0.47	0.42	0.36
Age (in years)	0.41	0.37	0.35
Education (in years)	0.19	0.14	0.12

Table 5: Coefficient estimates for the wealth regression models.

	Obs	MI all	Hot deck all	Overall var: (MI/HD)	Within var (MI/HD)
Intercept	-55.25 (-60.35, -50.15)	-55.33 (-60.14, -50.51)	-54.13 (-58.92, -49.34)	1.01	0.99
Fam.income	1.07 (1.01, 1.13)	1.1 (1.04, 1.16)	1.07 (1.01, 1.13)	1	1
Edu (in years)	0.53 (0.22, 0.85)	0.56 (0.26, 0.85)	0.57 (0.27, 0.87)	1	1
Non-Hisp black	-10.68 (-12.35, -9.01)	-11.57 (-13.12, -10.01)	-11.01 (-12.56, -9.47)	1.01	0.99
Other race/eth	-7.66 (-10.1, -5.23)	-7.47 (-9.77, -5.18)	-6.74 (-9.01, -4.47)	1.03	0.98
Age	0.95 (0.89, 1.01)	0.96 (0.9, 1.02)	0.96 (0.89, 1.02)	1	1
Never married	-0.01 (-2.07, 2.05)	0.18 (-1.8, 2.16)	0.01 (-1.93, 1.95)	1.04	0.99
Widowed	-8.55 (-12.25, -4.85)	-8.17 (-11.48, -4.85)	-8.1 (-11.41, -4.79)	1	0.99
Divorced	-10.66 (-12.95, -8.37)	-10.63 (-12.81, -8.44)	-10.82 (-12.97, -8.68)	1.04	0.99
Separated	-7.58 (-11.03, -4.13)	-7.55 (-10.9, -4.2)	-7.65 (-10.94, -4.35)	1.03	0.99

Note: The reference levels are non-hispanic white for race/ethnicity and married for the martial status.

Table 6: Mean values of total wealth (including home equities) and key wealth components across different methods and sources.

	Hot deck	MI	Hot deck–wt	MI–wt	SCF
Total net worth	202058	200383	316361	314129	470491
Business assets	29264	29564	48372	46963	110128
Checking/savings	19361	18404	29637	28725	47537
Stocks	32671	32845	56759	57438	67276
IRA/private annuities	33927	33636	52452	52144	60023
Net worth of vehicles	13444	13405	14301	14362	15002
Equity in primary residence	58408	58480	86427	86310	103932
Equity in real estate	18610	18461	28561	28547	52984
Other assets	8582	7903	10335	10033	24840
Other debts	12209	12316	10484	10393	11230

Table 7: Percentiles of total wealth (including home equities) and key wealth components across different methods and sources.

Percentiles	Hot deck	MI	Hotdeck-wt	MI-wt	SCF
Total net worth					
5th	-40000	-41141	-32000	-33000	-23600
10th	-15000	-15800	-9100	-9500	-4600
25th	0	0	2900	2500	6300
50th	20600	20000	54000	53650	64200
75th	138000	137000	270000	270000	226200
90th	488000	486000	790000	778800	794900
95th	927900	911534	1360000	1327683	1676300
Business assets					
90th	0	0	0	0	1001
95th	20000	20000	60000	67000	100082
Checking/savings					
10th	0	0	0	0	20
25th	0	0	20	20	580
50th	1200	1200	3000	3000	3803
75th	10000	10000	16000	17000	20016
90th	40000	38000	60000	60000	69357
95th	80000	80000	118000	116745	141916
Stocks					
90th	5000	4500	50000	50000	30025
95th	80000	80000	200000	200000	174142
IRA/private annuities					
75th	0	0	5000	5000	7006
90th	60000	60000	130000	130000	112092
95th	190000	197233	300000	300000	295241
Net worth of vehicles					
25th	1500	1500	2000	2000	3703
50th	7000	7000	8000	8000	9408
75th	18000	18000	20000	20000	18715
90th	30000	30000	35000	35000	32927
95th	45000	45000	50000	50000	44837
Equity in primary residence					
50th	0	0	15000	15000	26021
75th	65000	65000	105000	105000	116095
90th	175000	175000	250000	250000	265216
95th	275000	275000	355000	360000	417341
Equity in real estate					
90th	4000	5000	40000	40000	50041
95th	76000	75000	150000	150000	210172

Other assets					
75th	0	0	0	0	1501
90th	500	500	7000	6683	20016
95th	20000	20000	35000	35000	60049
Other debts					
50th	500	500	130	200	300
75th	10450	11000	8000	8500	8207
90th	34912	35000	30000	30000	28023
95th	60000	60000	51500	52000	52443

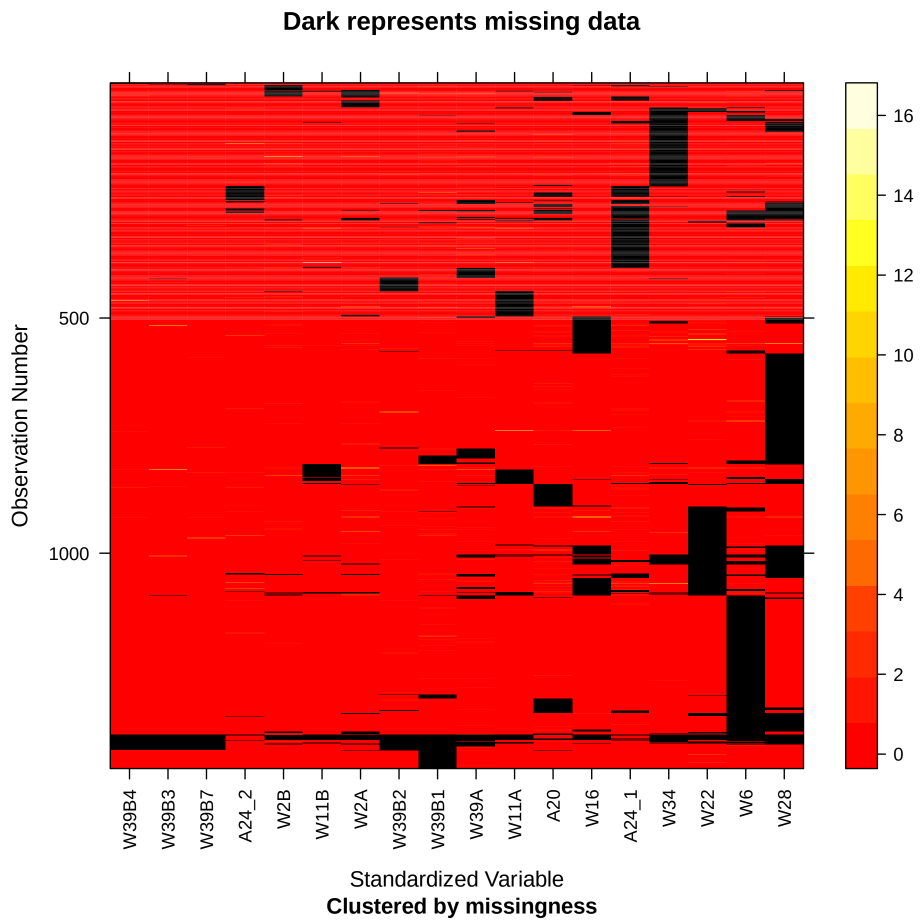

Figure 1: Missing patterns of 18 2013 PSID wealth components of the cases with missing wealth information. The plot is generated by the commands *mssing.data.frame()* and *image()* in the R package *mi*. The standardization process deducts the mean from a continuous variable and divides by twice the standard deviation of the observed values.

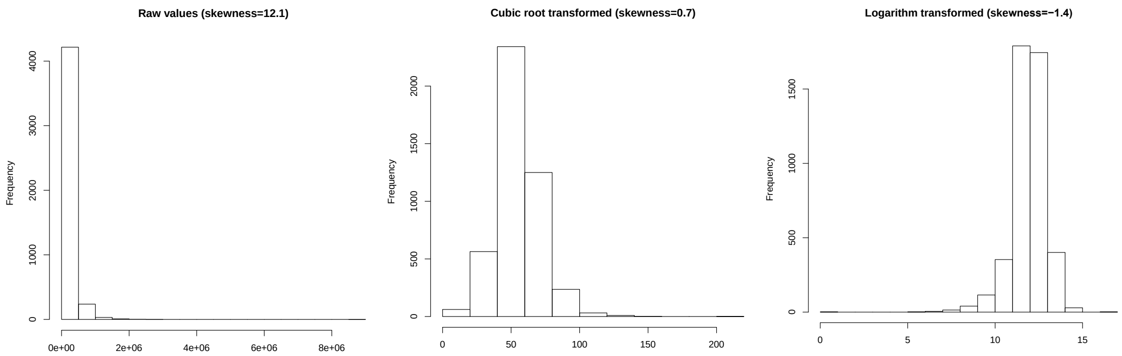

Figure 2: The distribution and skewness of home values before and after transformation.

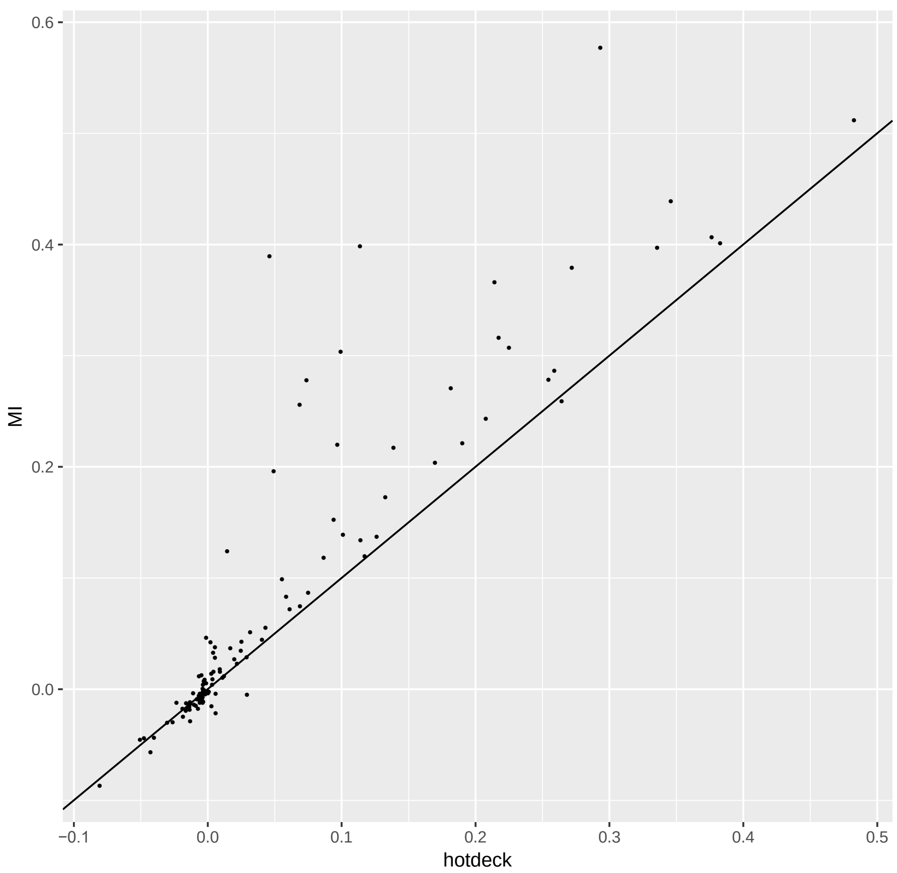

Figure 3: Pairwise correlation coefficients between 16 wealth components.

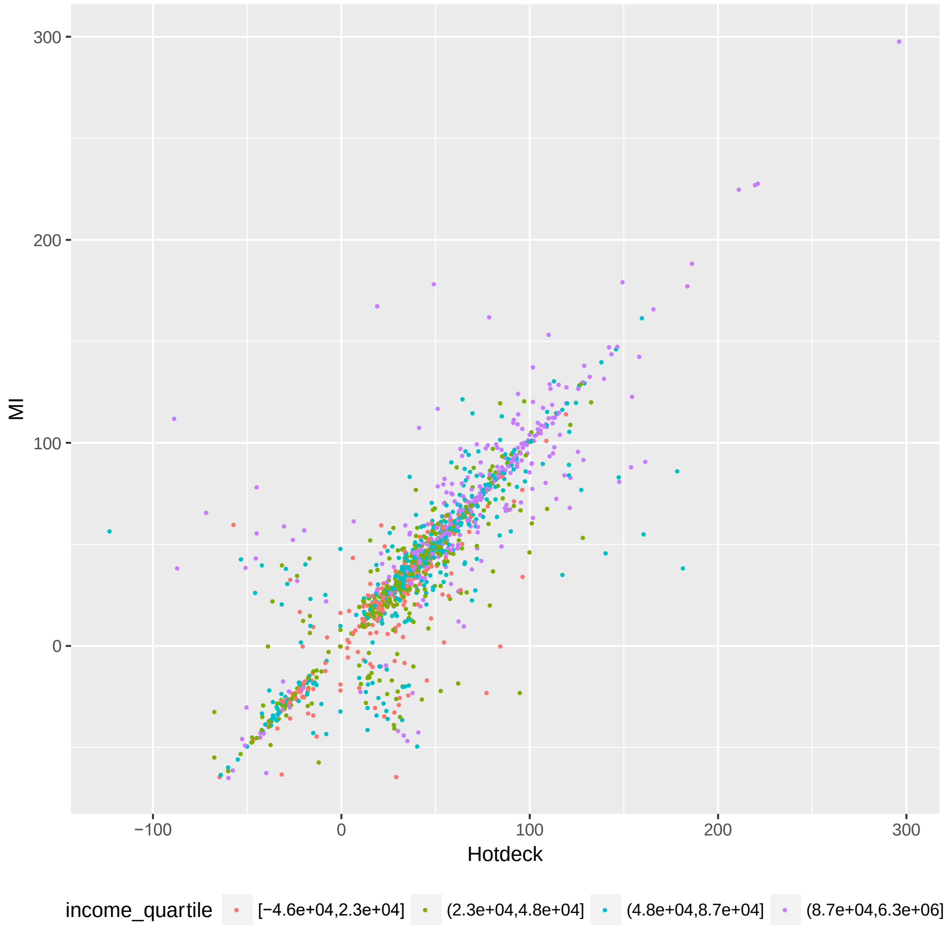

Figure 4: Comparison of imputed wealth values (cubic root transformed) from MI and hot deck imputation. The observations are marked in color grouped by income quartiles (in thousands): [-46, 23], (23, 48], (48, 87], and (87, 630]. Some wealth values are negative because of high debts and low assets.

Appendix 1: Names of variables included for wealth imputation in the 2013 PSID data (The labels are available in https://simba.isr.umich.edu/VS/s.aspx)

"AQCASE" "A4" "A8" "A19" "A20" "A21" "A22" "A23" "A23A_1" "A23B_1" "A24_1" "A28_1" "A23A_2" "A23B_2" "A24_2" "IMP_A20" "A20L" "A20H" "IMP_A19" "IMP_A23" "IMP_A24_1" "IMP_A28_1" "IMP_A24_2" "A24_1L" "A24_1H" "A24_2L" "A24_2H" "W1" "W1A" "W2A" "W2B" "W6" "W10" "W11A" "W11B" "W15" "W16" "W21" "W21A" "W22" "W27" "W28" "W33" "W34" "W38A" "W38BLEGL" "W38BMED" "W38BOTR" "W38BRELS" "W38BSTU" "W39A" "W39B1" "W39B2" "W39B3" "W39B4" "W39B7" "IMP_W11A" "W11AL" "W11AH" "IMP_W11B" "W11BL" "W11BH" "IMP_W28" "W28L" "W28H" "IMP_W2A" "W2AL" "W2AH" "IMP_W2B" "W2BL" "W2BH" "IMP_W16" "W16L" "W16H" "IMP_W6" "W6L" "W6H" "IMP_W34" "W34L" "W34H" "IMP_W22" "W22L" "W22H" "IMP_W39A" "W39AL" "W39AH" "IMP_W39B1" "W39B1L" "W39B1H" "IMP_W39B2" "W39B2L" "W39B2H" "IMP_W39B3" "W39B3L" "W39B3H" "IMP_W39B4" "W39B4L" "W39B4H" "IMP_W39B7" "W39B7L" "W39B7H" "IMP_W1" "IMP_W10" "IMP_W15" "IMP_W21" "IMP_W27" "IMP_W33" "IMP_W38A" "IMP_W38BSTU" "IMP_W38BMED" "IMP_W38BLEGL" "IMP_W38BRELS" "IMP_W38BOTR" "KEY_MISS" "NFU" "NKIDS" "NELDERS" "NFUHUCOLLEGE" "NOFUMCOLLEGE" "H49" "D1CKPT" "TOTHU" "GENMARSTAT" "IWLENMINS" "CCS" "G5" "G99" "G100" "G101" "G103" "G104" "G106" "G107" "G108" "G109" "G110" "G112" "G113" "G114" "WTRMETRO" "BNUM" "HBNUM" "WBNUM" "HWBIZ" "FARMY" "HDWG" "BNUS" "OVTM" "TIP" "COMS" "OTHINC" "PROF" "XTRA" "HDRENT" "HDDIV" "HDINT" "HDTRUST" "WFEARN" "WFRENT" "WFDIV" "WFINT" "WFTRUST" "TAXHW" "HDADC" "HDSSI" "HDOTRWELF" "HDVA" "HDPENS" "HDANN" "HDIRA" "HDUNEMP" "HDWRKCOMP" "HDCHSUP" "HDALIMONY" "HDHELPREL" "HDHELPNON" "HDMISC" "WFADC" "WFSSI" "WFOTRWELF" "WFPENS" "WFANN" "WFIRA" "WFUNEMP" "WFWRKCOMP" "WFCHSUP" "WFHELPREL" "WFHELPNON" "WFMISC" "OFLAB" "OFASSET" "OFADC" "OFSSI" "OFOTRWELF" "OFVA" "OFPENS" "OFUNEMP" "OFCHSUP" "OFMISC" "HDSSEC" "WFSSEC" "OFSSEC" "TOTFAMY" "FDHM13" "FDOUT13" "FDDEL13" "RENT13" "UTIL13" "TELINT13" "VEHLN13" "VEHPAY13" "VEHLS13" "AUTOIN13" "VEHADD13" "VEHREP13" "GAS13" "PARK13" "BUS13" "CAB13" "OTRAN13" "ED13" "CHILD13" "HOS13" "DOC13" "PRESCR13" "HINS13" "HHREP13" "FURN13" "CLOTH13" "TRIPS13" "OTHREC13" "FIPSTATE_DUM1" "FIPSTATE_DUM2" "FIPSTATE_DUM3" "FIPSTATE_DUM4" "FIPSTATE_DUM5" "FIPSTATE_DUM6" "FIPSTATE_DUM7" "FIPSTATE_DUM8" "FIPSTATE_DUM9" "FIPSTATE_DUM10" "FIPSTATE_DUM11" "FIPSTATE_DUM12" "FIPSTATE_DUM13" "FIPSTATE_DUM14" "FIPSTATE_DUM15" "FIPSTATE_DUM16" "FIPSTATE_DUM17"

"FIPSTATE_DUM18" "FIPSTATE_DUM19" "FIPSTATE_DUM20" "FIPSTATE_DUM21" "FIPSTATE_DUM22"

"FIPSTATE_DUM23" "FIPSTATE_DUM24" "FIPSTATE_DUM25" "FIPSTATE_DUM26" "FIPSTATE_DUM27"

"FIPSTATE_DUM28" "FIPSTATE_DUM29" "FIPSTATE_DUM30" "FIPSTATE_DUM31" "FIPSTATE_DUM32"

"FIPSTATE_DUM33" "FIPSTATE_DUM34" "FIPSTATE_DUM35" "FIPSTATE_DUM36" "FIPSTATE_DUM37"

"FIPSTATE_DUM38" "FIPSTATE_DUM39" "FIPSTATE_DUM40" "FIPSTATE_DUM41" "FIPSTATE_DUM42"

"FIPSTATE_DUM43" "FIPSTATE_DUM44" "FIPSTATE_DUM45" "FIPSTATE_DUM46" "FIPSTATE_DUM47"

"FIPSTATE_DUM48" "FIPSTATE_DUM49" "FIPSTATE_DUM50" "FIPSTATE_DUM51" "FIPSTATE_DUM52"

"FIPSTATE_DUM53" "FIPSTATE_DUM54" "FIPSTATE_DUM55" "FIPSTATE_DUM56" "FIPSTATE_DUM57"

"CHGMS_DUM1" "CHGMS_DUM2" "CHGMS_DUM3" "CHGMS_DUM4" "CHGMS_DUM5"

"CHGMS_DUM6" "CHGMS_DUM7" "CHGMS_DUM8" "FCC_DUM1" "FCC_DUM2" "FCC_DUM3"

"FCC_DUM4" "FCC_DUM5" "FCC_DUM6" "FCC_DUM7" "FCC_DUM8" "FCC_DUM9" "FCC_DUM10"

"WF_IND" "OF_IND" "SEX" "AGE" "ES" "BD1_FLAG" "BD2_FLAG" "GEOGMOBIL" "ENRLD"

"GD_ENRLD" "P0P70ACKPT" "P1" "P16" "P20" "P20L" "P20H" "H1HW" "H1AHW" "H1BHW" "H1CHW"

"H2HW" "H3HW" "H4HW" "H5AHWSTROK" "H5BHWHATTA" "H5CHWCORON" "H5DHWBLPRE"

"H5EHWASTHM" "H5FHWLUNG" "H5GHWDIABE" "H5HHWARTH" "H5IHWMEMLO" "H5JHWLEARN"

"H5KHWCANCE" "H5LHWNERVE" "H5MHWOTR" "H8HW" "H8ANITES" "COMPED" "KL39" "KL40_1"

"AVGHRS" "TOTOTHRS" "WKSILLOTR" "WKSILLSLF" "WKSVAC" "WKSSTRIKE" "WKSLAYOFF"

"WKSUNEMP" "WKSOOLF" "AGE_WF" "ES_WF" "BD1_FLAG_WF" "BD2_FLAG_WF" "GEOGMOBIL_WF"

"ENRLD_WF" "GD_ENRLD_WF" "P0P70ACKPT_WF" "P1_WF" "P16_WF" "P20_WF" "P20L_WF"

"P20H_WF" "H1HW_WF" "H1AHW_WF" "H1BHW_WF" "H1CHW_WF" "H2HW_WF" "H3HW_WF"

"H4HW_WF" "H5AHWSTROK_WF" "H5BHWHATTA_WF" "H5CHWCORON_WF" "H5DHWBLPRE_WF"

"H5EHWASTHM_WF" "H5FHWLUNG_WF" "H5GHWDIABE_WF" "H5HHWARTH_WF"

"H5IHWMEMLO_WF" "H5JHWLEARN_WF" "H5KHWCANCE_WF" "H5LHWNERVE_WF"

"H5MHWOTR_WF" "H8HW_WF" "H8ANITES_WF" "COMPED_WF" "KL39_WF" "KL40_1_WF"

"AVGHRS_WF" "TOTOTHRS_WF" "WKSILLOTR_WF" "WKSILLSLF_WF" "WKSVAC_WF"

"WKSSTRIKE_WF" "WKSLAYOFF_WF" "WKSUNEMP_WF" "WKSOOLF_WF"