Key words: Editing, Imputation, Panel data

Look Again: Editing and Imputation of SCF Panel Data

Arthur B. Kennickell*
Assistant Director
Board of Governors of the Federal Reserve System
Mail Stop 153
Washington, DC 20551
Arthur.Kennickell@frb.gov

Prepared for the 2011 Joint Statistical Meeting, Miami, Florida
August 3, 2011

Abstract

In 2009, a re-interview with participants in the 2007 Survey of Consumer Finances (SCF) was undertaken to provide information on the effects of the financial crisis on households. The panel questionnaire was designed to maximize comparability with the earlier data. The subject matter of the survey, wealth and related issues, is often considered sensitive or conceptually difficult. Consequently, editing and imputation of the data are very important considerations. Although the baseline data had already been edited and imputed cross sectionally, they were re-edited along with the new panel data. Similarly, the data for both waves of the survey were imputed jointly. This paper has two goals: to discuss the importance of the re-editing of the baseline data and to gauge the effects of the joint imputation of data from the two waves.

* Opinions presented in this paper are those of the author alone and they do not necessarily reflect the views of the Board of Governors of the Federal Reserve System or its staff. The author is particularly grateful to Brian Bucks, Jesse Bricker, Gerhard Fries, Kevin Moore, Traci Mach and other colleagues at the Federal Reserve; Barry Johnson, David Paris, Michael Parisi and others at Statistics of Income; and Deborah Cipriano, Katie Del Cielo, Catherine Haggerty, Ella Kemp, Shannon Nelson, Sandra Pitzer, Micah Sjoblom, Karen Veldman, Nina Walker, the interviewers for the SCF, and other central-office and field staff at NORC. Above all, the author is grateful to the survey respondents.

This paper addresses two practical questions about the statistical processing of panel data: How important is longitudinal editing? How important in imputation are data taken with a lag or a lead from the reference period of the particular data being imputed? One can answer these questions by taking a principled approach based on theory. That is, absent unobserved or uncorrectable biases of observation, one would generally want to use the maximum amount of information possible for both types of processing (e.g., see Little and Su [1989]). If one is short of the ideal level of resources or the ideal models are not available, however, it would be useful to know the degree of trade-off involved in using simpler approaches, such as purely cross-sectional editing and imputation. The author is only aware of two earlier attempts to investigate such questions, Kennickell and McManus [1994] and Frick and Grabka [2004].

The 2007–2009 panel of the Survey of Consumer Finances (SCF) offers compelling raw material for an empirically based analysis of the two questions. The survey deals with the details of households' finances and it uses a single instrument to interview a sample that ranges from the very poor to the extremely wealthy. As discussed further in the body of this paper, the interaction of the complexity and sensitivity of the subject matter in the SCF with variations in respondents' knowledge or willingness to cooperate fully led to a need for substantial editing and imputation.

The structure of the paper is as follows. The first section provides an overview of the design of the survey. The second section discusses the editing practices in the SCF and provides a picture of the degree of additional intervention in the baseline 2007 data necessitated by additional obtained in 2009. The third section discusses the FRITZ system used for imputation of the SCF and it presents evidence on the effects of inclusion of longitudinal information as conditioning variables in the imputations. A final section concludes and points to future research.

I. Background on the SCF

The SCF is usually executed as a triennial cross-sectional survey and the most recent such survey completed in 2009 was the 2007 SCF.[1] The economic turmoil following the subsequent onset of the financial crisis placed a high premium on gaining insights into the effects on the circumstances and behavior of households. In April 2009, the Federal Reserve Board

[1] See Bucks *et al.* [2009] for an overview of the content of the 2007 SCF.

authorized a panel survey to be undertaken with the 2007 survey participants. By July, the survey was in the field. The interval between authorization and execution was very short for such a complex undertaking. It is owing to the very high level of commitment of everyone involved, both at the Federal Reserve Board and NORC, that is was possible to construct technical materials and field procedures to deliver panel data that would enable a reliable comparison with the earlier cross-sectional data.

The questionnaire used in the triennial SCF collects very detailed information on the finances, attitudes and behavior of households, along with a variety of contextual information useful in analysis of the primary data. This instrument is highly structured to avoid double counting and to maximize reporting of various items in the appropriate places. One important cost off this structured approach to collecting such detailed information is the length of time required for an interview. In 2007, the typical interview required 75 to 90 minutes, but some very complex interviews required several hours and multiple sessions.

Both because it was feared that previous SCF participants would react negatively to another long interview and because time was very limited to prepare for the panel survey, it was decided to proceed with a maximally comparable, but substantially shorter, questionnaire. The panel instrument retains all of the high-level framing and sequencing of the questions in the baseline cross section, but generally far less detail was covered. The result was that the typical interview length dropped to between 45 and 60 minutes, and the right skewness of interview length was greatly limited.

Wealth in the U.S. is highly skewed—about about two-thirds of all household net worth is held by the wealthiest 10 percent and about half of that is owned by the wealthiest 1 percent. Thus, reliable overall analysis of wealth and of assets that are disproportionately held by wealthy households, such as bonds, requires sufficient representation of the wealthiest households. At the same time, the need to understand financial behavior more broadly distributed across the population necessitates the use of a sample with broad coverage. To meet both of these goals, the SCF employs a dual-frame design, including an area-probability sample and a list sample. The area-probability sample provides broad national coverage and a sample of households selected with equal probability (see Tourangeau et al. [1993]). For the 2007 SCF, about two-thirds of the survey participants derived from that sample. The 2007 list sample was selected using a model applied to a set of statistical records derived from individual income tax returns by

the Statistics of Income (SOI) Division of the Internal Revenue Service. The model was used to rank taxpayers in seven strata ordered by estimated wealth (see Kennickell [2001]). Observations with higher levels of predicted wealth were sampled at a higher rate. Weights were used to combine the two samples.

In the 2007 baseline survey, the respondent was the economically dominant single individual or the financially more knowledgeable member of the economically dominant couple; in some cases (particularly for cases with very wealthy or very ill respondents, a proxy was used to answer for the respondent). The great majority of the questionnaire focused on the "primary economic unit," which included all people in the household who were economically interdependent with the respondent and/or his or her spouse or partner.

Ideally, the 2009 panel interview would have been conducted with exactly the same person as in 2007 and the reference unit would have been the 2007 household. Unfortunately, even over the approximately two years between the baseline and panel interviews, some households underwent considerable change: couples got divorced, widowed or married, and other family members came or went. To make the field procedures for the panel feasible, the scope of the survey was limited to following at most one household for each 2007 household. The reference unit for the panel was defined as follows:

1. If the 2007 respondent was alive and not living permanently outside the U.S, the target household in 2009 was the one that contained that respondent.
2. Otherwise, if the 2007 respondent was either deceased or living permanently outside the U.S. and if the 2007 respondent had a spouse or partner who was a part of the PEU as defined in the 2007 survey, the target household in 2009 was the one that contained the 2007 spouse or partner of the 2007 respondent, if that spouse or partner was still living and residing permanently in the U.S.
3. Otherwise (where (a) the 2007 respondent was either deceased or living permanently outside the U.S. in 2009 and (b) either (i) there was no spouse or partner who was a part of the 2007 primary economic unit or (ii) there was such a spouse or partner but that person was either deceased or living permanently outside the U.S.), the case was considered to be out of scope for the 2009 survey.

There were 4,422 households that participated in the 2007 SCF. The response rate for the area-probability sample was about 70 percent and that for the list sample ranged from about 50

percent for the least wealthy stratum to about 10 percent for the wealthiest. Despite the fact that there had been no indication at the time of the 2007 survey that there might be a follow-up interview and the fact that names were not available for all sample members, it was possible to determine whether virtually every case was in scope or not and to contact the overwhelming majority of those who were in scope. The response rate in 2009 conditional on 2007 participation was about 89 percent across a wide range of economic and demographic groups (see Kennickell [2010]). The final number of participants in 2009 was 3,862.

II. Data editing

The SCF questionnaire can be challenging for both the respondents and the interviewers. Some of the topics covered are difficult for some people to understand, it is sometimes difficult to locate relevant records, and the subject matter is most often seen as very private. Moreover, language appears to have an inherent fluidity that allows for what amounts to a distribution of meaning over any of the elements of the question, and some of the meanings may differ in important ways from the intended meaning. This problem is exacerbated in the case of the SCF because the questionnaire is long and sometimes classifications turn on points that require particular attention from the respondent and the interviewer. The questionnaire must accommodate a range of financial conditions ranging from desperately poor to astonishingly wealthy. Despite best efforts to avoid overly technical language and multiple qualifiers, to streamline the questioning, to prepare interviewers for the potentially most difficult parts of the questionnaire, and to include automated checking of parts of the data during the interview, there are seemingly inevitable problems.

One way the SCF copes with problems is by encouraging interviewers and respondents to make comments during the interview to clarify unusual responses and responses where there is a question about the appropriateness of the classification of the information. In addition, interviewers are required to complete a debriefing questionnaire for each case, and as a part of that process, to review all of the comments recorded during the interview and make necessary clarifications. Taking all of the commentary and any verbatim records of responses from respondents together with a list of potential anomalies identified by a computer program run on the raw data, the project staff reviews all of the data for each survey case (see Kennickell [2007]). Where there are clear deviations from the intended survey protocol,

adjustments are made in an effort to maximize the clarity with which the situation of each respondent is represented.[2] To maximize comparability over time, edit decisions follow a substantial body of "case law" derived from the earlier history of the survey.

The 2009 panel introduced a new element into this SCF routine. Having a second observation on the participants allowed the possibility of detecting errors in the 2007 data that were previously obscured. This possibility was greatest when there were logical constraints on outcomes, such as the ordering of events in time. To assist in the detection of such problems, the battery of post-field diagnostics applied to the data was modified for the panel specifically to search for potential inconsistencies between the two waves of data.[3]

To facilitate the data editing, the editing staff was given a formatted view of the survey responses that interwove the 2009 data for each case with comparable values computed from one of more variables in the less aggregated 2007 data. All supplemental commentary and diagnostics from both years were available, as were the more detailed 2007 micro data. Where the review revealed a resolvable inconsistency that potentially affected the baseline and panel data, a strenuous effort was made to minimize changes to the baseline data.

Based on this longitudinal editing, about 700 observations were determined to have errors that required at least some adjustment to the original 2007 data. Many of these edits were relatively small. By far, the greatest numbers of inconsistencies affecting dollar-denominated variables were associated with job-related pensions, often for the plans of the spouse or partner of the survey respondent. Prior SCF experience and other research indicate that many people have a poor understanding of their job-related retirement plans; indeed, recognition of this problem had motivated a simplification of the SCF pension questions for the 2004 survey. Although that redesign appears to have improved data quality, the pension questions can only be hoped to collect what people understand well enough to recall or interpret from their records. The particular errors mostly clustered into two groups: plans that were reported in 2009 as

[2] In the longer run, the survey digests patterns of errors in order to develop appropriate future refinements in the questionnaire, interviewer training and materials aimed at respondents.

[3] In principle, some such detection could have been made a part of the CAPI program for the 2009 panel interview—for use in structured edits or in framing the new answers. Largely out of concern that respondents might be made suspicious by the potential direct or indirect recall of their earlier data, such information was restricted to 2007 housing tenure and ownership of a private business in which the household had an active management role. Changes in these two situations were critical for the subsequent analysis of the data. Even had there been a desire to introduce 2007 data into the interview, it would have required more complicated programming that time allowed, because of the presence of missing or partially missing (range) data.

having existed from substantially before the 2007 interview but that were not reported in 2007, and plans that were inconsistently reported as either a defined-benefit (annuity) plan or an account-type plan such as a 401(k) account. In many instances, the implication of such inconsistencies was that entire sequences of questions needed to be set to missing in the 2007 data.

The next most frequent source of recognizable inconsistency was in the reporting of mortgage loans on the primary residence. The available data along with the commentary from the interviewers suggest that some respondents had incorrectly reported the original loan they took out when purchasing the home, rather the loan in force at the time of the interview. In general, these later loans would have been taken out to refinance the original loan. Most often, these inconsistencies affected some of the terms of the loans, but not the amount outstanding on them. Resolution of a wide scattering of other problems led to other changes in the 2007 data, sometimes with the implication that some values were set to missing.

Because the full 2007 data set has not yet been fully re-imputed as of the time of writing this paper, it is not yet possible to make a meaningful quantitative assessment of the effect of changes to the data due to editing on the estimates of key characteristics of that cross section. In the results reported later in this paper that use wealth data, the largest source of potential discrepancy, pension accounts, are excluded.

A serious question facing researchers who produce panel data is the extent to which baseline or other previously released data should be re-issued with corrections when errors are detected. The SCF data for various years have been corrected a number of times as inconsistencies or other errors have been resolved. Based on this precedent, it was decided to apply the edits implemented for the panel version of the 2007 data in more extended form to the full 2007 data set. The argument for doing so, however, is not unambiguous. Even though the earlier changes to a given set of cross-sectional data have occurred at the level of individual cases, every case in the survey had the possibility of being examined and corrected. In contrast, the panel edits based on the 2007–2009 panel data can only apply to cases that responded in both years. One mitigating argument is that the high response rate in the panel leaves only a small fraction of the in-scope cases unexamined in light of new data. For the cases out-of-scope in 2009—largely cases lost to death—the great majority of the types of edits are much less likely to apply to their data. The original cross-sectional data will be maintained separately.

III. Imputation

Missing data are a substantial problem in the SCF. The great majority of cases have at least some missing information. Although the fraction of missing information is normally quite small, overall at least some data are missing for a very wide of variables. In general, variables indicating ownership or receipt of a given item, attitudinal variables and variables indicating demographic characteristics are rarely missing, but dollar denominated-variables can have more substantial missing data rates. Because dollar-denominated variables are an essential element of the SCF, the survey instrument allows for a variety of ways of reporting such information, including a facility for capturing range information (see Kennickell [1997]). Although an exact point answer is the best response to a question about a dollar amount for most analytical purposes, the markets for the item may not be well enough developed that the respondent could look up the value in their records or a market report, even in principle. For example, it is generally not possible to know the value of a personal business until an effort is made to sell it. A respondent may not know the exact value of a potentially more precisely defined item, and may be unable to unwilling to consult records. In addition, some respondents are simply wary of giving too precise an answer. In such cases where respondents are unable or unwilling to provide an exact answer, they can often be persuaded to give a range.

Tables 1a and 1b report on the response status of a selection of dollar-denominated variables for the cases in the 2007 and 2009 waves of the SCF panel. For purposes of these tables, data are counted as missing if the respondent was unable or unwilling to give either a single non-missing response or a range, or the value was set to missing during the data editing. Aside from the case of pension assets given in the table, the rate of completely missing data tends to be fairly low. Experience in the SCF indicates that many people have little knowledge about their retirement assets, even in the case of straightforward accounts like a 401(k). The rate of completely missing information on pension accounts in 2007 also reflects, in part, the result of editing that caused a substantial number of such values to be set to missing, as noted earlier.

As is very clear from the tables, range information is very important for the SCF. In the great majority of cases, the ranges are not overly wide. Although it is possible for respondents to give a one-side range (e.g., “more than $1 million”), only a very small fraction of range responses are of this type. It should be noted that to a degree, the higher rate of range reports in 2007 than in 2009 is an artifact of the construction of the aggregated 2007 variables for the panel

data set. For example, if a respondent in 2007 had four savings accounts, gave a complete response about the amount in each of the first three but only had a vague idea of the small amount in the fourth one, the aggregated amount constructed for the panel data set would appear as a range; although there is a conceptually parallel possibility for 2009 in that a respondent might go through the exercise of adding up all such accounts and give a very narrow range, it seems far more likely that the respondent would simply approximate the small remaining amount.

Table 1a: Missing data, 2007 SCF; percent of cases where question was not known to be inapplicable.

Item	Unweighted %			Weighted %			Weighted % of total	
	Good	Range	Missing	Good	Range	Missing	Range	Missing
Monthly rent	94.4	5.4	0.1	94.4	5.6	0.0	5.3	0.1
Value main house	90.7	8.8	0.5	91.4	8.0	0.6	5.9	0.5
Mortgage outstanding	86.0	9.5	4.5	83.5	11.0	5.4	8.4	2.5
Checking balances	81.4	15.1	3.5	81.4	14.6	4.0	10.0	3.9
Mutual fund holdings	69.9	22.3	7.8	73.2	20.1	6.7	16.0	11.6
Stock holdings	74.2	17.8	8.0	73.5	19.6	6.9	23.5	10.9
Pension account balances	60.7	22.9	16.5	56.9	25.9	17.2	12.8	25.5
Credit card balances	94.4	4.8	0.8	93.3	6.0	0.7	9.1	0.5
Total wages	77.9	20.5	1.6	77.6	20.3	2.1	11.0	1.7
Self-employment/farm income	77.0	21.1	2.0	73.3	22.7	3.9	32.2	2.8
Other business income	74.0	23.8	2.2	77.6	16.7	5.6	33.9	3.0
Pensions and Social Security income	80.3	16.4	3.4	78.2	16.8	5.1	19.1	5.5

Note: Mutual fund and stock holdings include such funds held outside retirement accounts or other managed accounts; pension account balances refer to the amount in job-related retirement accounts for the family head; all other variables refer to totals for the family.

.

Table 1b: Missing data, 2009 SCF; percent of cases where question was not known to be inapplicable.

Item	Unweighted			Weighted			Weighted % of total	
	Good	Range	Missing	Good	Range	Missing	Range	Missing
Monthly rent	95.8	2.9	1.3	95.6	3.0	1.3	5.5	1.0
Value main house	90.5	8.3	1.1	88.6	9.9	1.5	5.7	0.5
Mortgage outstanding	87.2	6.1	6.6	84.9	7.4	7.7	5.8	3.2
Checking balances	90.1	6.7	3.0	89.2	7.6	3.0	5.0	3.6
Mutual fund holdings	80.7	11.8	7.3	79.5	13.6	6.9	21.6	5.0
Stock holdings	80.3	10.9	8.4	79.1	12.6	8.3	14.7	5.9
Pension account balances	69.9	14.2	15.8	66.1	16.0	17.8	14.7	14.7
Credit card balances	96.7	2.6	0.6	95.5	3.8	0.6	6.2	1.1
Total wages	86.0	11.2	2.7	85.1	12.5	2.3	7.9	2.8
Self-employment/farm income	81.5	12.8	5.5	81.7	14.3	3.8	20.1	8.4
Other business income	84.8	6.9	8.0	88.9	6.3	4.8	11.9	1.6
Pensions and Social Security income	83.7	12.6	3.8	83.5	13.3	3.2	12.5	4.5

Note: See note to table 1a.

In the SCF, all missing data are multiply imputed using an implementation of the FRITZ system originally developed for the survey (see Kennickell [1998]). The program includes a series of models for individual variables and it incorporates the possibility of constraining outcomes using ranges, institutional or logical constraints, and other such prior information. The most commonly used techniques within this application amount to randomized predictions of linear regressions tailored to the patterns of available (non-missing or already imputed) data for each observation. The program is run over a number of iterations until the results are stable. For each iteration, the covariance input for the regression-like models is computed using the complete data matrix from the prior iteration.

The original 2007 survey data were fully imputed using an updated version of the system of programs that has been used for every SCF beginning with the 1989 survey. Following recommendations from Little and Su [1989], an entirely new implementation of the FRITZ system was created to impute the 2009 data and to re-impute the 2007 data, where these imputations are conditioned on information from both years.

Although re-imputing previously released data has a strong attraction on theoretical grounds, such an approach does have complications beyond the considerable work of constructing a new set of models. In most panel surveys, not every case present in the baseline has complete observations beyond that period, so it would not be possible to construct such jointly conditioned imputations of baseline data for all cases. In addition, there are potential costs to both the data creators and the data users in dealing with results that might differ between the two sets of data. Moreover, despite the compelling arguments of Little and Su, there would undoubtedly be users, particularly in the social sciences, who would find it suspect to use future information for imputing past variables.

For the SCF panel, generally only the aggregated variables for 2007 comparable to those collected in 2009 were re-imputed, not the more detailed variables of the original data set. The aggregated variables will be released together with the 2009 data and with warnings of the potential for differences from the original detailed 2007 data (and the re-edited version discussed earlier in this paper). For the sake of completeness, data for 2007 will also be included in the panel data set for participants in 2007 who were deemed out of scope or who otherwise did not participate in 2009; aggregated variables for those cases will be constructed directly from the

imputed cross-sectional data. Analysts will be able to link the panel data to the original data file, but caution will be urged.

To understand the potential for different analytical results as a consequence of the imputation strategy for the panel, it is useful to consider alternatives that allow the possibility of identifying some of the potential differences. This work may have use not only for the narrow purposes of the SCF but also for others considering how to deal with imputation in panel data.

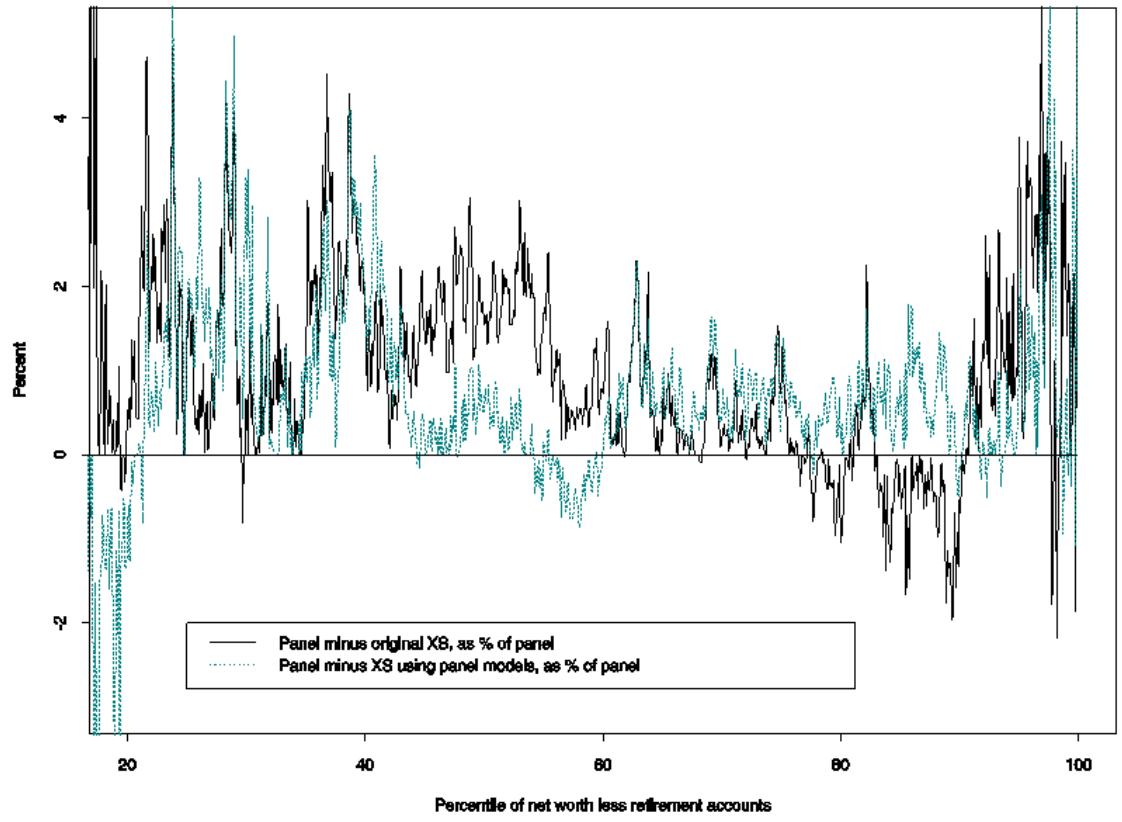
First, we compare three approaches to imputation of the 2007 data: (a) using the full panel imputation system to impute both years of data, (b) using the 2007 cross-sectional imputations, and (c) using the panel models for the 2007 data with all covariates based on 2009 data set to missing. Alternative "c" allows one to gauge the effects of omitting some of the detailed variables available for the original cross-sectional models. The equations in the box below help to show the differences one might expect to see in these models.[4]

Informally speaking, if a variable Y_t is missing at random conditional on X_t in a linear framework, then the cross-sectional models (options "b" and "c," equation (1)) would be expected to yield unbiased predictions of Y_t. If adding one-period leads of X and Y to the model (option "a," equation (2)) similarly does not induce bias, then the overall difference in the two approaches would be the relative size of the variances of ε and ζ; if there is additional information for Y_t in X_{t+1} and Y_{t+1}, then the variance of ε would be greater than the variance of ζ. If, however, Y_t is not fully missing at random conditional on X_t, but is either missing at random or "closer" to that condition when X_{t+1} and Y_{t+1} are added to the model, then we would expect differences in the distributions of the imputations under the two approaches.

(1) Cross-sectional model:	$Y_t=X_t\beta+\varepsilon_t$
(2) Panel model:	$Y_t=X_t\theta+X_{t+1}\lambda+Y_{t+1}\gamma+\zeta_t$

[4] For the experimental imputations using the panel models, the imputation system was not iterated in the usual way. Full iteration would have required far more resources than were available. Instead, the final jointly imputed panel data were treated as the lagged iteration for purposes of computing the necessary moment matrices for further use of the panel models in the experiments reported here. A small test done using only the unimputed data as input for these calculations (as is the case for a first iteration of the imputations) indicates that the decision to use the panel data as input had little effect on the conclusions of the comparisons across experiments.

Figure 1: Relative quantile-difference plots for total 2007 family net worth less designated retirement assets.



To explore the differences in distributions under these approaches, figure 1 shows a set of relative quantile-difference plots (the percent difference in the quantile values of two distributions) for options "b" and "c" relative to option "a" for total family net worth less any type of explicitly retirement-related account (IRA, Keogh, 401(k), etc.).[5] The horizontal axis shows the quantile points of the distribution and the vertical axis measures the percent difference in the quantile values of the distributions. The horizontal axis is truncated below about the 20th percentile, because below that point the percent swings in the estimates are exaggerated by small nominal movements that are very large percentage amounts—particularly for amounts close to zero in the baseline distribution. The figure indicates that the panel models overall tend to predict slightly higher wealth across the distributions. However, the differences are generally quite small—under about 2 percent, a range that would not be significant if imputation and sampling error were considered.

[5] In all cases, the reference group is the population that participated in the 2009 panel interview.

Figure 2: Relative quantile-difference plots for total 2006 family income.

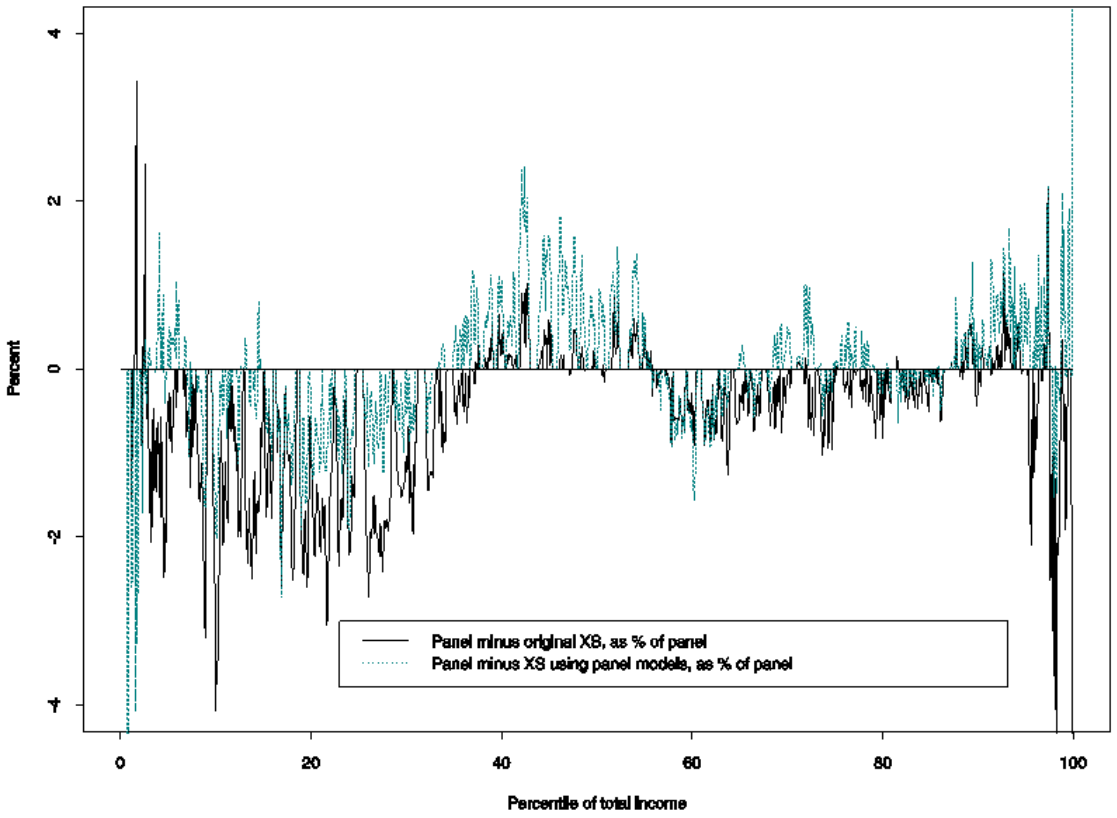


Figure 2 shows comparable comparisons for total family for the calendar year preceding the baseline survey, 2006. The figure shows that the data including the results of the full panel models predicts slightly less income in approximately the bottom third of the distribution, slightly more in the center of the distribution and slightly less in the upper part of the distribution. As in the case of figure 1, the differences would not be significant when considered against sampling and imputation error.

The variables considered in each of the two figures apply to all members of the population and both are economically important variables that show a high degree of dispersion and that require an important degree of imputation. The slight differences in the plots suggest that at least differences in the univariate distributions under the difference imputation schemes were not much affected by the choice among imputation strategies. Perhaps because of the relatively small number of multiple imputations (five), there is no clear ordering among the relative imputation variances either.

Probably the most compelling motivation for imputing the 2007 and 2009 SCF data jointly is to preserve, to the extent possible, inter-period correlations for the sake of longitudinal analysis of the data, which is the primary driver for collecting the 2009 data. One way of probing the effects of the imputation on such analysis is to perform regressions mixing data from the two waves of the survey.

Table 2 shows a series of regressions of families' total financial assets in 2007 on their income in various categories collected in the 2009 survey. Because there are often distinct age patterns in financial assets, the model also includes a quadratic in the age of the household head. The models are run on data created under four scenarios. In the first three, the 2009 data are imputed using the panel models and those models condition on data from both years; the 2007 data are imputed following the three strategies described earlier. In the fourth scenario, both the 2007 and 2009 data are imputed using the panel models, but conditioning only on data from the survey year of the data imputed.

Table 2: Coefficients and standard errors due to imputation for a regression of the logarithm of 2009 financial assets (for those having any) on variables defined using the 2007 data, for various patterns of imputation of 2007 and 2009 data.

	Imputation strategy															
Models for 2007:	Panel models				Original 2007 models				Panel models				Panel models			
Conditioned on:	2007 and 2009				2007				2007				2007			
Models for 2009:	Panel models				Panel models				Panel models				Panel models			
Conditioned on:	2007 and 2009				2007 and 2009				2007 and 2009				2009			
Variable	All cases		Cases with some missing data		All cases		Cases with some missing data		All cases		Cases with some missing data		All cases		Cases with some imputation	
	Coeff.	*SEI*	Coeff.	*SEI*	Coeff.	*SEI*	Coeff.	*SEI*	Coeff.	*SEI*	Coeff.	*SEI*	Coeff.	*SEI*	Coeff.	*SEI*
Intercept	-0.03	*0.00*	-0.02	*0.01*	-0.03	*0.00*	-0.02	*0.00*	-0.02	*0.00*	-0.02	*0.00*	-0.03	*0.00*	-0.02	*0.01*
R_AGE07	-4.27	*0.04*	-3.12	*0.07*	-4.17	*0.04*	-2.92	*0.08*	-4.31	*0.07*	-3.23	*0.16*	-4.27	*0.05*	-3.20	*0.10*
R_AGESQ07	0.44	*0.00*	0.33	*0.01*	0.43	*0.00*	0.31	*0.01*	0.45	*0.01*	0.33	*0.01*	0.44	*0.00*	0.33	*0.01*
D_LABINC07	0.49	*0.02*	0.43	*0.03*	0.47	*0.01*	0.42	*0.05*	0.49	*0.01*	0.42	*0.05*	0.48	*0.02*	0.38	*0.06*
L_LABINC07	0.05	*0.00*	0.04	*0.00*	0.05	*0.00*	0.05	*0.00*	0.05	*0.00*	0.05	*0.00*	0.05	*0.00*	0.05	*0.00*
D_BUSINC07	-1.23	*0.02*	-1.38	*0.05*	-1.30	*0.02*	-1.60	*0.06*	-1.21	*0.01*	-1.34	*0.07*	-1.19	*0.02*	-1.30	*0.05*
L_BUSINC07	0.52	*0.00*	0.54	*0.00*	0.52	*0.00*	0.56	*0.01*	0.51	*0.00*	0.53	*0.01*	0.51	*0.00*	0.53	*0.01*
D_CAPINC&YR	-4.22	*0.07*	-3.47	*0.30*	-4.40	*0.11*	-3.28	*0.35*	-4.46	*0.15*	-3.51	*0.59*	-4.44	*0.15*	-3.56	*0.42*
L_CAPINC07	0.43	*0.01*	0.35	*0.04*	0.45	*0.01*	0.33	*0.04*	0.46	*0.02*	0.35	*0.06*	0.45	*0.02*	0.35	*0.04*
D_PENINC07	5.94	*0.05*	6.37	*0.15*	5.92	*0.06*	6.26	*0.12*	5.96	*0.04*	6.45	*0.04*	5.91	*0.05*	6.32	*0.12*
L_PENINC07	1.91	*0.01*	1.78	*0.02*	1.91	*0.01*	1.79	*0.02*	1.91	*0.01*	1.79	*0.01*	1.92	*0.00*	1.80	*0.00*
RMSE	0.06	*0.00*	0.05	*0.01*	0.06	*0.00*	0.05	*0.01*	0.06	*0.00*	0.05	*0.00*	0.06	*0.00*	0.06	*0.01*
N	3673		1085		3673		1085		3673		1085		3673		1085	

Notes: R_AGE07 is the 2007 age of the household head, R_AGESQ07 is R_AGE07**2/100; "D" variables are dummy variables for ownership, where 0=the family does not have the item and 1=the converse; "L" variables are the logarithm of the maximum of 1 and the income values; LABINC07 is total family wages for 2008; BUSINC07 is total family income from a business, self-employment or real estate in 2008; CAPINC07 is total dividends and taxable and nontaxable interest income in 2008; PENINC07 is total income from annuities, defined-benefit pensions and Social Security; RMSE is the root mean squared error of the regression. SEI is the standard error with respect to imputation only.

The estimates based on all cases vary slightly across the different versions of the data, and sometimes the differences are larger than the standard error with respect to imputation. But none of the estimates are significantly different when the standard errors of the coefficients are used (not shown). Although nearly 30 percent of the observations originally had a missing value for at least one variable included in the regression, it might still be that the "real" data dominate the outcome. To examine this possibility, the same models were also run using only the cases

that originally had at least some missing data. Owing largely to the smaller sample size, there is somewhat more variability in the estimates across the different imputation scenarios. But in no case are the differences significant. Although it is not feasible to examine regressions of even all of the major variable clusters in this detail, the selected alternatives considered did not deliver a substantively different outcome.

If join imputation were important in the SCF, it seems very likely that it would show up in comparison of estimates of changes in families' net worth and income between the two waves of the panel under different imputation strategies. Table 3 shows selected quantiles of the distribution of these changes under four strategies for net worth less designated retirement assets and for total income. The first three take options "a," "b" and "c" discussed earlier for imputing the 2007 data and the 2009 data are imputed conditional on both 2007 and 2009 values of variables. The fourth strategy uses the panel models for both years, but includes only data specific to each survey year.

Table 3: Quantiles and mean of change in net worth less designated retirement assets and change in total family income for the previous year, 2007-2009; thousands of dollars.

	Imputation strategy			
Models for 2007:	Panel models	Original 2007 models	Panel models	Panel models
Conditioned on:	2007 and 2009	2007	2007	2007
Models for 2009:	Panel models	Panel models	Panel models	Panel models
Conditioned on:	2007 and 2009	2007 and 2009	2007 and 2009	2009
Variable Statistic				
Net worth less designated retirement assets				
25th percentile	-70.5	-72.0	-72.0	-72.0
Median	-8.2	-8.5	-8.6	-8.5
75th percentile	14.4	14.9	14.7	15.3
Mean	-84.2	-90.8	-89.8	-87.4
Total income				
25th percentile	-9.5	-9.4	-9.5	-9.6
Median	1.3	1.4	1.3	1.3
75th percentile	12.9	12.9	12.7	12.7
Mean	-5.3	-4.5	-5.5	-5.1

If there is an important relationship between variables in the two years, then joint modeling of the two survey waves should be expected to yield less noisy measures of change. In the case of net worth less designated pension assets, the values of the measures of the distribution of change using the jointly imputed estimates are a bit lower than the corresponding values under all of the alternative imputation strategies.[6] The difference is most pronounced for the imputation strategy that uses the original cross-sectional data for 2007 and the jointly imputed data for 2009. For the most obvious alternative, conditioning on only own-year data for each survey, however, the differences are generally smallest. If we consider total income instead

[6] Note that a lower estimate does not necessarily imply lower measurement error here.

of net worth, the differences between the estimates based on full panel imputation and all of the alternatives are negligible, except for the case of the mean value based on the difference between the full panel imputation of the 2009 data and the original cross-sectional data for 2007.

IV. Conclusions and future research

Overall, the estimates examined in this paper appear fairly robust to method of imputation considered for the SCF panel data. There are a variety of potential explanations for this surprising result. An ever-present possibility is an error in the complicated system that was developed for the imputations and the experimental variations on that system. This possibility seems small, given the independence of the experimental variations and the degree of checking of the output. The lack of strong variation in results may be a function of the particular statistics reported—surely, there are variables or combinations of variables that would show noticeable differences across the variations. Another possibility is the use of the jointly imputed panel data to compute the moment matrices for the experimental imputations using the panel models, but at least limited testing using the unimputed data as input suggest that this is not an important factor.

As noted earlier in this paper, the SCF obtains a great deal of information in the form of ranges. It seems not implausible that the ranges might have been sufficient to "anchor" the distribution of the data closely enough that any variations across methods would be swamped by random variation. One way to test this possibility would be to impute the data under all the alternatives treating the range information as entirely missing. This exercise is beyond the scope of this paper.

Finally, it may be that there are relationships in the data that are specific to the period of the surveys. The 2007 wave of the panel was undertaken before the largest part of the downturn in the value of real estate, businesses and financial assets had taken place; to a degree, the previous run-up in values may have created a greater contrast across households. The 2009 wave showed both large changes from the baseline overall and great heterogeneity in the outcomes across families.

A further test with a different period of the SCF would be useful. Although there may be a regular SCF panel in the future, there are not definite plans as of this time. An alternative might be to consider another panel survey of wealth that was in existence before the recent

financial crisis, such as the Encuesta Financiera de las Familias (EFF) conducted by the Banco de España (Bover [2010]).

Bibliography

Bover, Olympia [2010] "The Spanish survey of household finances (EFF): description and methods of the 2008 wave," Documento Ocasional de próxima publicación, Banco de España, http://www.bde.es/webbde/es/estadis/eff/eff2008_be1210.pdf.

Bucks, Brian K., Arthur B. Kennickell, Traci L. Mach and Kevin B. Moore [2009] "Changes in U.S. Family Finances from 2004 to 2007: Evidence from the Survey of Consumer Finances," Federal Reserve Bulletin, v. 95, pp. A1–A55.

Bucks, Brian, Jesse Bricker, Arthur Kennickell, Kevin Moore and Traci Mach, [2010] "Surveying the Aftermath of the Storm: Changes in Family Finances from 2007 to 2009," Finance and Economics Discussion Series (2011-17, March), Board of Governors of the Federal Reserve System, http://www.federalreserve.gov/pubs/feds/2011/201117/201117pap.pdf.

Frick, Joachim R. and Markus M. Grabka [2004] "Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income distribution," Discussion Paper 376 (April), DWI German Institute for Economic Research.

Kennickell, Arthur B. [2010] "Try, Try Again: Response and Nonresponse in the 2009 SCF Panel," *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, Vancouver, British Columbia, Canada.

Kennickell, Arthur B. [2007] "Look and Listen, But Don't Stop: Interviewers and Data Quality in the 2007 SCF," *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, Vancouver, Salt Lake City, Utah.

Kennickell, Arthur B. [1998] "Multiple Imputation in the Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, Dallas, TX.

Kennickell, Arthur B. [1997] "Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances," http://www.federalreserve.gov/pubs/oss/oss2/papers/rangepap0197.pdf.

Kennickell, Arthur B. and Douglas A. McManus [1994] "Multiple Imputation of the 1983 and 1989 Waves of the Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, Toronto, Ontario.

Little, R.J.A. and Su, H.-L. [1989] "Item Non-Response in Panel Surveys" in Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. P. (eds.), *Panel Surveys*, John Wiley, New York: 400-425.

Tourangeau, Roger, Robert A. Johnson, Jiahe Qian, Hee-Choon Shin, and Martin R. Frankel [1993] "Selection of NORC's 1990 National Sample," working paper, National Opinion Research Center at the University of Chicago, Chicago, IL.