FDZ

Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ·Methodenreport

Methodological aspects of labour market data

02/2012
EN

Multiple imputation of household income in the first wave of PASS

Ursula Jaenichen,
Joseph W. Sakshaug

Bundesagentur für Arbeit

Multiple imputation of household income in the first wave of PASS

Ursula Jaenichen, Joseph W. Sakshaug (Institute for Employment Research (IAB))

Contents

Tables and figures

Abstract

The report summarizes the results of a project aimed at the completion of household income in the first wave of the “Panel Arbeitsmarkt und Soziale Sicherung" (PASS) by using multiple imputation routines. The imputation approach chosen is an iterative procedure combining individual information on respondents and non-respondents with household level information. The report discusses the various steps of the imputation procedure and demonstrates some quality aspects of the imputed data.

Zusammenfassung

Der Bericht fasst die Ergebnisse eines Projekts zusammen, das auf die Vervollständigung des Haushaltseinkommens in der ersten Welle des “Panel Arbeitsmarkt und Soziale Sicherung" (PASS) mittels multipler Imputation zielt. Der gewählte Imputationsansatz ist eine iterative Prozedur, in der Informationen für befragte und nicht befragte Personen mit Informationen auf Haushaltsebene kombiniert werden. Der Bericht diskutiert die einzelnen Schritte der Imputation und demonstriert einige Qualitätsaspekte der imputierten Daten.

Keywords: Multiple imputation, item non-response, household income, combined household and individual data, Panel Arbeitsmarkt und Soziale Sicherung (PASS)

We are grateful to Stefan Bender, Jörg Drechsler, Martina Huber, Frauke Kreuter, Johannes Ludsteck and Mark Trappmann for very helpful comments and discussion and to Daniel Gebhardt, Antje Kirchner and Gerrit Müller for help with the data.

1 Introduction

This report summarizes the results of a project exploring the possibilities to raise the quality of the information on household income contained in the first wave of the “Panel Arbeitsmarkt und Soziale Sicherung" (PASS) by using multiple imputation routines. PASS (see e.g. Hartmann et al. 2008) is an ambitious panel survey providing unique information on German households with and without receipt of unemployment benefit Alg2.

Item non-response in surveys may lead to biased statements about population characteristics if the process of missingness is not arbitrary. In addition, basing analyses on complete cases only is equal to losing the information contained in incomplete observations. Thus, the resulting parameter estimates will be less precise than necessary. Applying multiple imputation, missing data are substituted with draws from a predictive distribution. Correct standard errors for parameter estimates are obtained by combining the estimates of several imputed data sets.

Multiple imputation can be used to deal with non-response for all variables contained in a data set. The results discussed here focus on the completion of household income in the first wave of PASS. To a certain degree this is due to the complexity of the PASS data, for which household and individual records – both containing numerous variables of potential relevance for status and income – can be combined. Therefore, the report gives an example of how item non-response in PASS may be adressed, with household income being a variable of central interest to many researchers.

The imputation approach chosen is an iterative procedure combining individual information on respondents and non-respondents with household level information. Individual earnings are imputed conditional on observed or estimated labor market status and household income. Thereafter, household income is imputed using the sum of the individual incomes of household members as a predictor in the imputation model.

2 PASS

PASS is a longitudinal household survey aimed at providing data for “labour market, welfare state and poverty research in Germany” (Trappmann 2011, 10). For the first wave, there are two main samples. The BA sample is drawn randomly from “Bedarfsgemeinschaften”[1] from administrative data on benefit (Alg2) receipt, the Microm sample is a household sample of the German population in which households with lower social status are overrepresented. Interviews are carried out on the household level and on the individual level, in the first wave the household questionnaire asks some basic questions on each household member, including non-respondents. There are special questionnaires for elder (>65 years) persons and questionnaires translated to other languages. Individual weights as well as household weights are made available to take the sampling design into account.

[1] Bedarfsgemeinschaften can consist of some or all household members sharing a common budget. They are the base for receipt of benefit Alg2.

The number of households interviewed in the first wave is 12794, with 6804 households belonging to the BA sample and 5990 households belonging to the Microm sample. In total, 18954 persons were interviewed, 9386 from the BA sample and 9568 persons from the Microm sample. The average response rate on the household level is about 31 percent, and higher in the BA sample (35 percent) than in the Microm sample (27 percent). Within households, the response rate is reported to be about 85 percent and nearly equal for the two samples.

The variety of topics addressed in the PASS interviews is remarkable (Beste et al. 2011). The household questionnaire asks for subjects like e.g. living conditions, housing and housing costs, household income and child care. The individual questionnaire contains blocks of items on e.g. socio-economic background, attitudes, work and unemployment, leisure activities, social integration, pensions and health related issues.

3 Method and Software

Imputation is a common procedure used to adjust for item nonresponse in surveys. It is especially common to apply imputation to income variables as these items tend to elicit the highest rates of missing data. The pattern of missing income data is believed to be nonrandom as persons with especially low or high incomes are less likely to report their incomes due to privacy concerns. Thus, if the income data are analyzed without correcting for item nonresponse, the resulting inferences may be biased.

An advantage of multiple imputation Rubin (1987, 1996) is that it can help correct nonresponse bias and it provides data users with a complete rectangular data set that can be analyzed using standard statistical software. In addition, if the imputations are performed by the survey agency and released to data users, then different data users will be able to obtain the same inferences when performing the same analyses on the imputed data. This is a significant advantage over letting data users apply their own missing data adjustment, which could yield conflicting results between users.

Although single imputation is often used to adjust for item nonresponse, multiple imputation is the preferred and most principled approach. By generating multiple imputations for each missing value, it is possible to account for the uncertainty of imputed values. The imputed values consist of draws from a predictive distribution. By drawing multiple values from this distribution it is possible to obtain an estimate of the between-imputation variance that reflects the uncertainty of the imputed values. There is no way to account for the uncertainty of a single imputed value, which is why single imputation yields standard errors that are too small, confidence intervals that are too narrow, and p-values that are too significant. This is the main reason why multiple imputation is the preferred imputation approach.

Despite its many advantages, there are significant challenges to using multiple imputation in large surveys. The main challenge is specifying a joint distribution of all of the variables to be imputed. A joint distribution is needed to preserve the associations between all of the variables. Specifying such a distribution is an extremely challenging task in large surveys where there are hundreds of variables representing different distributional forms (e.g., continuous, binary, mixed, etc.). A practical alternative to joint modeling is called sequential regression multiple imputation (or chained equations; see Raghunathan et al. 2001, Oudshoorn et.al. 1999). Instead of modeling all variables in a single joint model, the chained equations approach models each variable one-at-a-time, and conditions on all other variables to preserve the associations between the variables. This approach is advantageous because it doesn't require a fully joint model, rather it uses a univariate model for each variable. Each variable can be modeled separately using a model that is appropriate for each variable type, such as linear regression for continuous data, logistic regression for binary data, and so on. The corresponding models are then used to impute the missing values. The chained equations procedure is the preferred imputation approach in large surveys, and is built-in to several statistical software packages, including R, Stata, SAS, or IVEware.

The imputation software *ice* used here is available as a Stata ado-file (Royston 2005a, 2005b, 2007, 2009). It offers comfortable options to include different variable types (continuous variables, binary indicators, ordered and non-ordered categorical variables). Equations for every variable can (and sometimes must[2]) be specified separately. The option "conditional" allows to restrict the estimation to subgroups, important for dealing with filtered variable structures. Semi-continuous variables, which are truncated at 0 from below, can be treated by two-step-modeling: in the first step a binary variable indicating whether the variable takes a positive value or otherwise is modeled and imputed for cases with missing values for the semi-continuous variable. The second step consists of a regression estimation and the imputation of the continuous variable part, conditional on the binary indicator being equal to one (Drechsler 2011, Seaman/White 2008, Ragunathan et al. 2001, Yu et al. 2007).

Ice contains a special feature for the automatic treatment of "perfect prediction", which may occur when the dependent variable is categorical. In this case, the dependent categorical variable completely "separates" an independent variable or a combination of independent variables, meaning that the categories of the dependent variable are perfectly corresponding with different value ranges of the independent variable(s). This makes estimation and especially the determination of standard errors impossible. The solution implemented in *ice* consists in augmenting the data with a few observations, thus allowing for the estimation of the model without biasing the estimation results (Royston 2007, White/Royston 2010). However, the occurence of perfect prediction might also be the result of errors in the specification of the model and this should be checked before relying on the automatical procedure.[3]

An option which is highly important for the treatment of the income variable is the possibility to deal with interval censoring or "bracketed" variables (Drechsler 2011, Royston 2007, Schenker et al. 2006). As the information on household income as well as the information on individual earnings is asked first as an exact amount and thereafter in intervals getting stepwisely finer, the censored regression model seems to be a good way to make use of all the observed information.

[2] As a default, *ice* includes all variables in the equations.

[3] When the models described in this report were developed, it was often possible to avoid perfect prediction by dropping variables from single equations or by regrouping categories.

Statistical analyses on the imputed data can be performed applying Rubin's rules (Rubin 1987, 1996): the same analysis is done for each of the *m* imputed data sets and the results are thereafter combined to obtain final estimates. The final parameter estimate is given by the mean of the *m* single estimates; the variance can be calculated as follows:

$$T_m = \overline{U}_m + \frac{m+1}{m} B_m$$

Here, $\overline{U}_m$ is the average within-imputation variance of the estimated parameter, while B_m is the variance across imputations.

The *mim*-program implemented in Stata provides the automatic calculation of parameters and standard errors using the formula above (Royston et al. 2009).[4]

4 Imputation based on household variables

The complexity of the PASS data represents a challenge for anybody who does not only want to pursue his/her individual research interests, but wants to provide a dataset which may be exploited by other researchers interested in the analysis of the PASS data.

There are a couple of advantages in constructing an imputation model which is based exclusively on the variables contained in the household questionnaire:

- the number of variables is manageable
- the variable "net household income" nicely aggregates different sources of income of the household members
- there is no multi-level structure

While these advantages facilitate the imputation and possibly raise its transparency, the drawbacks are obvious:

- the information on individual incomes of household members is not used
- information on the determinants of individual and household incomes is neglected

[4] While the program also allows to combine descriptive summary statistics, in this report it is only used to obtain regression coefficients and standard errrors.

4.1 Information on household income in PASS

In the household questionnaire, net household income is first asked as an exact amount and thereafter – if the exact amount is not given - in intervals getting finer in several steps.

The variable HEK0600 contained in the first wave gets positive values only if the exact amount of household income had been revealed in the interview. This was true for 88.5 % or 11318 households out of a total number of 12794 households interviewed. The variable hhincome is provided by the PASS people and combines interval information and exact information. The interval information either represents interval mid points or empirical median values (if there is only upper or lower bound).

Table 1: PASS information on net household income

Variable	Obs	Mean	Std. Dev.
HEK0600	11318	1591.826	1333.461
hhincome	12423	1633.882	1817.021

The combined income information is missing only for 2.9 % of the households. As has already been noted, the program *ice* contains an option for the imputation of interval censored variables. For households with known bounds on income, the imputation model will predict values lying within these bounds. Thus, if not observed, the imputation procedure will provide predictions of the exact amount of household income for all other households in the data set.

4.2 Simple household model

From the remaining household variables, the variables in figure 1 were chosen as relevant for the imputation model:

Figure 1: Variables in the household model

- ✓ sum of payments the household receives from other households (semi-continuous)
- ✓ sum of payments the household gives to other households (semi-continuous)
- ✓ household weight
- ✓ BA/Microm (binary)
- ✓ household language (categorical)
- ✓ federal state (categorical)
- ✓ household size
- ✓ number of BG's
- ✓ household composition (categorical)
- ✓ housing type (categorical)
- ✓ 5-point scale for condition of dwelling house (ordered)
- ✓ 5-point scale for condition of residential area (ordered)
- ✓ square meters (continuous)
- ✓ per-person housing costs (separate interval regressions for owners/non-owners, conditional on housing type)

- ✓ subsidies to housing cost (semi-continuous, conditional on filter variables)
- ✓ costs for child care (semi-continuous)
- ✓ transfer receipt (3 binary indicators for different types of transfer, conditional on filter variables)
- ✓ household debts (ordered)
- ✓ mother or father reduced working time because of child care responsibilities (2 binary indicators, conditional on filter variables)
- ✓ index of deprivation (count, reweighted)
- ✓ 11-point-scale for actual living condition of household
- ✓ 11-point-scale for future (expected) living condition of household

Because of the structure of the questionnaire, the variables are centered around housing and general living conditions, leaving aside individual determinants of household income like education or employment status.

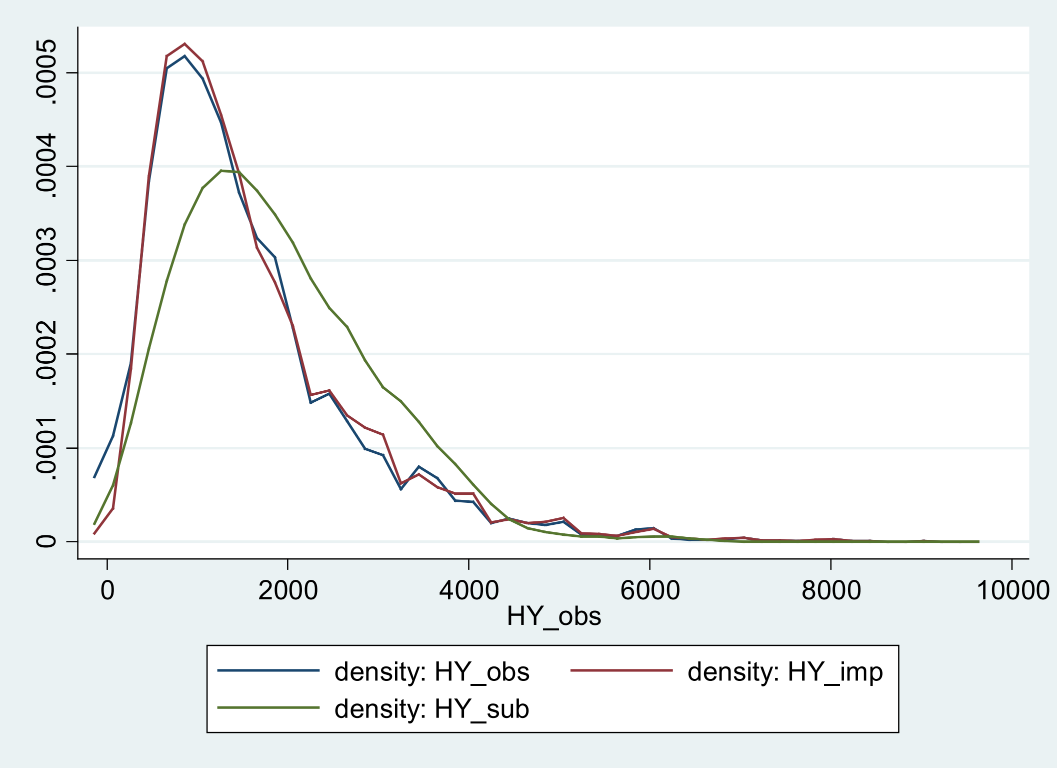
4.3 Results of the simple household model

The result of the imputation procedure are *m* completed data sets in which missings for each variable have been substituted by plausible values based on the posterior predictive distribution resulting from the estimated equation. A number of *m=5* data sets seems to be reasonable (Drechsler 2011). Here, we confine the inspection of the completed data to the distribution of the income variable in one version of the imputed data sets.

The graph shows density functions for

- the observed household income HY_obs (combining exact and interval information),
- the completed variable HY_imp from one of the imputed data sets and the
- the variable HY_sub containing imputed income values for those 373 households without any observed income information.

Figure 2: Kernel densities for different versions of household income



The observed and the completed income variables show very similar distributions. The density for the completed income variable *HY_imp* has a slightly higher peak than the PASS generated variable *HY_obs*. In contrast, the density for those incomes imputed for households without income information is clearly more to the right than the other two other densities.

Table 2: Deciles of the distribution of different versions of household income

	dec1	**dec2**	**dec3**	**dec4**	**dec5**	**dec6**	**dec7**	**dec8**	**dec9**
HY_obs	541	700	900	1100	1300	1509	1840	2300	3000
HY_imp	540	700	900	1100	1300	1520	1900	2400	3088
HY_sub	650	951	1233	1466	1700	2032	2346	2786	3346

This result is again demonstrated in table 2 which lists deciles of the distribution of the different income variables. It shows that the imputed income values HY_sub for households without observed income information are higher than the observed and completed income values for all households along the whole distribution of household income. Assuming that the imputation model correctly maps the determination of household income, this implies that higher income households are more inclined not to reveal their net incomes.

5 Imputation based on person and household variables

In order to make use of the information contained in the PASS person records as well, the final project step combines a model of individual labor income with the model of household income already discussed. The idea is that household income should be more or less equal to the sum of the individual incomes of household members and that this relationship should be exploited.

However, there are (at least) two issues presenting complications of this basic idea. The first arises because of the coexistence of respondents and non-respondents within households. Thus, for the households contained in PASS, there is a non-negligible share of persons for whom no interview was realized (in addition, children under age 15 are not supposed to be interviewed). Discounting non-respondents within households would bias the estimated household incomes downward, if estimated as the sum of individual incomes.[5]

The second issue concerns the various income categories contained in the PASS data in combination with filtering and/or different questionnaires for subgroups. The differentiation of these categories like gross/net wages, wage from extra jobs, wage from mini-jobs[6], pensions, unemployment benefits, etc. get quickly messy when trying to find out which persons possibly should/could have information on which type of income.

The solution found here is as follows (see Schenker et al. 2006 for a similar approach):

- For non-respondents within households, there is some basic information (age, gender) obtained in the household interviews. The imputation model for individual wages is estimated for a joint sample of respondents and non-respondents. This implies that for non-respondents labor market status and wage income are predicted on the basis of a very small number of observed variables.

- Assuming that wages or labor income are the most important source of household income, the imputation of missing individual earnings has been focussed on this income category. For other income categories, ad-hoc-procedures were used to deal with missingness. There are other income categories like pensions or benefits worth to be looked at more closely.

- The imputation model for individual earnings is performed conditional on labor market status or job type. This is a newly generated variable which describes the type of job. Unemployment and inactivity is contained as an extra category.

[5] The importance of dealing with non-respondents within households is highlighted in Frick et al. for the German Socio-Economic Panel Study (SOEP).

[6] In Germany, mini-jobs are low-income jobs exempted from social security contributions to a large degree.

The imputation can broadly be outlined as follows:

- perform a preliminary imputation of household income based on variables of the household questionnaire only - generate one complete data version
- repeat the following three steps *m* (chosen to be 5) times
  1) perform imputation of individual wages for respondents and non-respondents, using some household information of data completed in the preceding step and again generate one complete data version
  2) add up observed and imputed incomes within households
  3) perform final imputation of household income

5.1 Labor market status and job type

The imputation model for persons predicts wages conditional on labor market status. This makes sense, because there should be no positive wages if persons are inactive or unemployed. Wages will be influenced by qualification and working time, furthermore, for some job types, they are largely determined by institutional arrangements.

Because of the central role of labor market status, it is discussed before turning to the results on individual incomes. A new variable has been created in order to distinguish between working and non-working persons as well between job categories. This categorization of labor market status tries to remove some of the heterogeneity contained in the determination of labor incomes, hopefully leading to a more precise prediction of earnings.

Figure 3: Categories of labor market status / job type

(1) working time >=16, net wage>=400, no extra job
(2) working time >=16, net wage>=400, extra job(s)
(3) working time >=16, net wage<400, with or without extra job(s)
(4) working time <16, net wage>=400, with or without extra job(s)
(5) working time <16, net wage<400, without extra job(s)
(6) mini-job
(7) apprenticeship
(8) unemployed/out of the labor force

Earnings are imputed only if a person's labor market status is observed or estimated to be in the categories (1) to (5). These categories include self-employed persons as well, as PASS directly asks for the monthly amount these persons would get out of their enterprise for consumption. Persons with more than one job are asked to report their earnings from extra jobs. The earnings variable used in the imputation model is the sum of net earnings in the main job and in extra jobs. Earnings from mini-jobs and earnings for apprentices are not imputed. For both groups, earnings determination is seen as predominantly dictated by institutional/legal norms, thus different from the categories (1) – (5).

Table 3: Labor market status / type of job in observed data and in the 5 imputed data sets

	observed data	**imputed data**				
Status	0	1	2	3	4	5
hours 16+, Ynet 400+, no extra jobs	19.4	26.4	26.6	26.3	26.6	26.4
hours 16+, Ynet 400+, extra jobs	1.4	1.8	1.8	1.8	1.8	1.8
hours 16+, Ynet < 400	0.8	1.1	1.1	1.1	1.2	1.1
hours<16, Ynet 400+	0.9	1.2	1.2	1.2	1.2	1.0
hours<16, Ynet < 400	1.0	1.4	1.4	1.4	1.4	1.3
Mini-Jobs	7.5	10.0	10.0	10.4	9.7	10.0
Apprentices	2.0	2.4	2.4	2.3	2.4	2.2
Unemployed/not in labor force	43.7	55.8	55.6	55.6	55.8	56.0
Missing	23.3	0.0	0.0	0.0	0.0	0.0
Total	**24019**	**24019**	**24019**	**24019**	**24019**	**24019**

Table 3 compares the distribution of labor market status in the observed and in the imputed data (the imputation model is described below). In the observed data (dat=0), for 23% of the persons no labor market status is observed – this comprises both interviewed persons and persons without an interview. The distributions of labor market status in the imputed data are very similar across the five versions. After the imputation, all non-missing categories have increased somewhat, with the strongest increases (in percentage points) in the two largest groups: the share of workers in the first category (working time of 16 h or more, earnings of more than 400 €, no extra-job) has risen from 19 to 26 percent, and the share of unemployed or inactive people has risen from 44 to 56 percent.

5.2 Imputation of individual earnings

Like household income, earnings are in PASS first asked as an exact amount and thereafter in intervals getting stepwisely finer. The earnings model for persons contains the household variables income and household size and an indicator of whether the household makes part of the BA sample or the Microm sample. Individual sample weight is included, for non-respondents this weight has been substituted by the quotient of household weight and household size. Other individual variables roughly summarize a person's socioeconomic background. Experimenting with different specifications, it turned out to be difficult to add more individual variables to the model without getting warnings from *ice*.[7]

[7] The final model specification resulted in a warning of perfect prediction associated with "female".

Figure 4: Variables in the individual model

- ✓ wage income/plus income from extra jobs - interval regression
- ✓ household income (from 1st/previous imputation)
- ✓ household size
- ✓ individual sample weight
- ✓ BA/Microm (binary)
- ✓ migration background (binary)
- ✓ West/East (binary)
- ✓ age (cubic root)
- ✓ female (binary)
- ✓ married (binary)
- ✓ labor market status (categorical)
- ✓ educational level (categorical)
- ✓ occupational status (categorical)

The reason probably can be found in highly correlated variables like own and parent's schooling attainment together with clustered sample structures like the rather small population share of foreigners or migrants in East Germany.

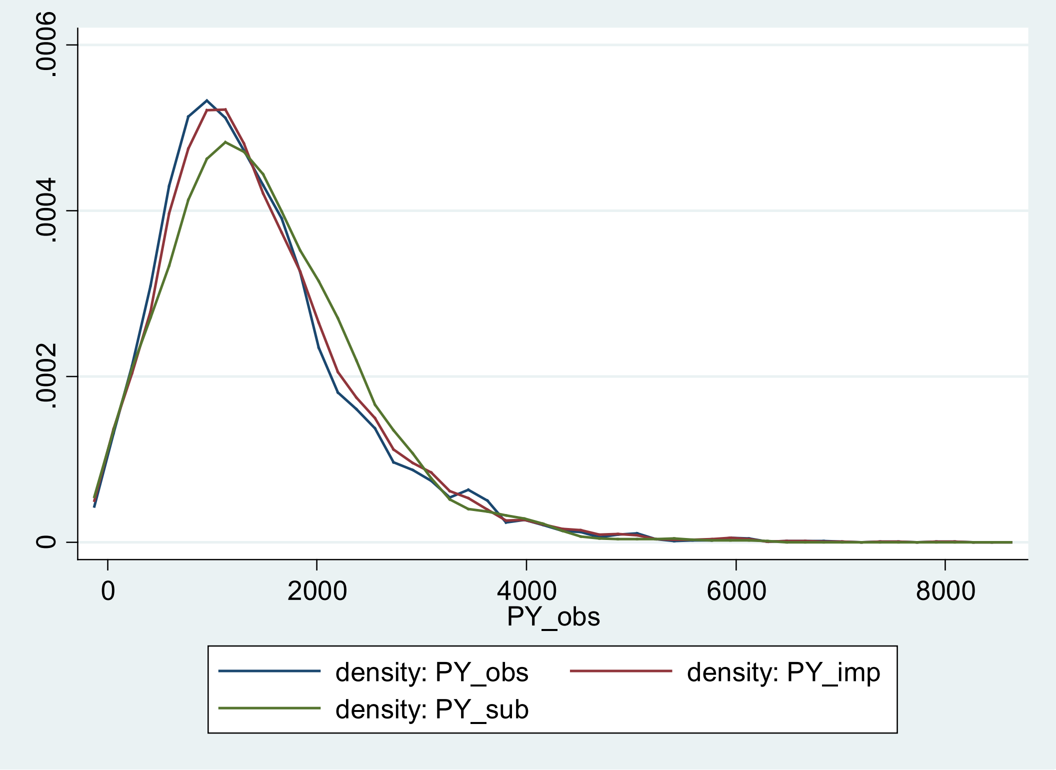
5.3 Results of the individual earnings' model

Figure 5 shows the density functions of

- a PASS generated variable *PY_obs* combining exact and coarsened information on individual earnings,
- an imputed earnings variable *PY_imp* from one run of the individual imputation model for persons with observed or predicted labor market status in categories 1-5,
- imputed earnings *PY_sub* for a subgroup of persons who did not report any information (including non-respondents without an interview).

While observed earnings and imputed earnings for all persons in status categories 1-5 have very similar distributions, the distribution of imputed earnings for the subgroup of persons without reported earnings lies somewhat more to the right than the other two densities.

Figure 5: Densities for observed and imputed individual earnings



The deciles listed in table 4 allow for a more precise statement of the differences in observed and imputed earnings distributions. As before, the imputed variable is chosen from one imputation run.

Table 4: Deciles of the distribution of different versions of individual incomes

	dec1	**dec2**	**dec3**	**dec4**	**dec5**	**dec6**	**dec7**	**dec8**	**dec9**
PY_obs	500	700	900	1100	1279	1500	1750	2100	2700
PY_imp	480	709	918	1100	1300	1500	1800	2139	2791
PY_sub	425	745	995	1173	1369	1610	1829	2181	2673

Again, one can see that the distributions for observed and completed data are very similar, the deciles for the imputed earnings variable PY_imp being with one exception slightly higher than in the distribution for the observed earnings variable. Looking only at imputed earnings of persons without any observed earnings information, only the first and the ninth decile are lower than in the data for all persons. The median earnings for these persons are 1369 Euros which is 90 Euros above the median in the observed data. Thus, similar to the household model results, potential earnings seem to be higher for non-interviewed persons and for persons not reporting their earnings.

5.4 Imputation of individual earnings: a regression

To get an impression of the plausibility of the imputation results a simple regression model for individual earnings is run once on the observed data and thereafter on the imputed data using Stata's *mim* program. The idea is that severe implausibilities in the imputed data should show up in this comparison. The results are presented in table 5. For some variables like "age", "age squared" or "medium school" the coefficients estimated on the imputed data are somewhat different from those estimated for the observed data, but all in all the results from the two regressions are very similar. Some of the estimated standard errors in the imputed data regression are slightly larger than those in the observed data regression. The interpretation is that using the imputed data has not generally led to more precision, but is counteracted by the uncertainty contained in the imputation itself. This is might be due to the fact that the imputation model used is rather crude and that the individual data contain a high number of non-respondents for whom predictions will have a greater variance between across imputations.

Table 5: Regression results for observed and imputed data, dependent variable log earnings

	Observed Data			**Imputed Data**		
	Coeff.	Std.Err.	p-value	Coeff.	Std.Err.	p-value
age	0.0678	0.0071	0.000	0.0550	0.0045	0.000
age sqared	-0.0723	0.0084	0.000	-0.0493	0.0047	0.000
migration background/non-German	-0.1338	0.0295	0.000	-0.1376	0.0347	0.001
female	-0.5652	0.0238	0.000	-0.5517	0.0272	0.000
married	0.1472	0.0266	0.000	0.1386	0.0278	0.000
educational level (ref. high school)						
no degree	-0.2230	0.0810	0.006	-0.2492	0.0778	0.002
medium school	0.1188	0.0300	0.000	0.0755	0.0340	0.036
prof. college entry level	0.2829	0.0497	0.000	0.2717	0.0464	0.000
college entry level	0.3867	0.0323	0.000	0.3819	0.0366	0.000
constant	5.6131	0.1435	0.000	5.7093	0.1109	0.000
obs	4699			7616-7712		
R^2	0.1749					

5.5 Summing up individual income

After each imputation, individual earnings are summed up within households. If available, earnings from other income categories are added as well. These are incomes from mini-jobs, wages of apprentices (Auszubildendenvergütung), unemployment benefit (alg1), pensions, parental leave allowance, and the household- or "Bedarfsgemeinschaft"-oriented unemployment benefits (alg2). Different procedures for dealing with missing values for these income categories were chosen in order to limit the modeling effort without wasting the available information or introducing additional biases.

- imputed: wages and income from extra job
- completed with mean value for observed cases:
  mini-job income, wage of apprentices
- completed with mean value of the population in the relevant age: unemployment benefit alg1, pensions
- added only if observed: parental leave allowance
- added once per household: amount of benefit receipt alg2[8]
  (variable completed in the first imputation using household data).

The sum of earnings from the main and eventual extra jobs is imputed conditional on labor market status falling into one of the categories 1 to 5 in the individual earnings model. For mini-jobbers and apprentices, missing values for earnings are substituted with the mean observed values in these categories. For status category 8, unemployed and inactive persons, unemployment benefits and pensions are important income sources. However, not everybody in the category will have earnings from these sources. To avoid explicit models of unemployment and pension receipt, for persons below age 65, the mean of unemployment benefits weighted with the observed probability of unemployment receipt in this category is substituted for persons without information on benefit receipt[9]. Analogously, for persons of age 65 or above, the mean observed pension weighted with the probability of pension receipt in this age-status combination is substituted if persons have missing information on pension receipt. Receipt of parental leave allowance is of minor importance in the observed data and thus added only if observed. The amount of benefit receipt alg2 is imputed in the first household imputation model and added here to the sum of individual incomes once per household.

As can be seen from table 6, on average and for the median value, the sum of individual incomes remains well below the household incomes. This pattern is still stronger in the observed data. Here, the sum of individual incomes simply adds up observed income compo-

[8] This assumes identity of household and Bedarfsgemeinschaft.

[9] For respondents, in a first step, the mean observed value of unemployment benefits alg1 and the mean observed value for pension is substituted for persons with a high probability of alg1 or pension receipt but without information on the amount of these income categories. The method of substituting mean values for certain income categories does not always seem to give plausible results, for example in cases in which the sum of observed individual incomes exactly equals observed household income. An alternative method left for future work would be to randomly allocate these income amounts among potential (as determined by status) recipients.

nents in the same way as has been described for imputed data. As non-respondents' income is neglected, the greater gap between the sum of individual incomes and the household incomes makes sense.

Trying to explain this gap, it can be noted that the gap is unevenly distributed across households: based on the last imputation, from a total of 12794 households, for 19 percent or 4492 households, the sum of individual earnings exceeded household net income. Non-response not perfectly dealt with in the imputation model and the summing up procedure might be one reason, as the size and the distribution of the gap changed when different possibilities to substitute income categories like alg1 and pensions for non-respondents were experimented with. Another reason for the gap might be the fact that certain income categories are not asked for in the PASS questionnaires and/or not included in the imputation and thus left incomplete. One example are 1-euro-jobs, for which the amount of earnings is not asked for and for which simple indicators are included in the household model equations. There are also some income categories like child benefits (Kindergeld) or subsidies to education (BAföG) which are included in the household model but ignored when summing up individual incomes. Still another reason might consist in misconceptions about the definition of income or a tendency to be more precise about some income categories than others.

Table 6: Median and mean of household incomes and sum of individual incomes after each person level imputation run

	Sum of individual incomes		**Household income**	
	Median	Mean	Median	Mean
Observed data	1000	1395	1300	1634
After imputation no.				
1	1100	1523	1300	1634
2	1100	1519	1300	1632
3	1103	1520	1300	1629
4	1100	1511	1300	1631
5	1100	1521	1300	1631

5.6 Final household model

For the final imputation of household income, the sum of individual earnings is used as an additional regressor in the imputation model. Other modifications to the first household model are indicators for the number of non-respondents in the household and the number of children under age 15. Further indicators of missingness of some earnings' component are binary indicators for 1-euro-jobs, for missing information on earnings from an extra job, for missing information on the amount of parental leave allowance. Receipt of benefit alg2, previously imputed within the household model, is now excluded (because included in the sum of individual incomes). The multinomial indicator of receipt of alg2 is transformed into a binary indicator of actual and former benefit receipt to enable estimation.

Figure 6 presents kernel densities of the PASS generated income variable HY_obs, the imputed income from one run of the final model, HY_imp and the imputed income only for households who did not report their income, neither the exact amount nor any information on income bounds. As a first impression, these densities do not look very different from those resulting from the model using household information only. The density for the imputed household income variable HY_imp is very close to the density for observed income HY_obs. The density for households who did not inform about their income lies visibly more to the right than the other 2 densities.

Figure 6: Kernel densities for different versions of household income, final model results

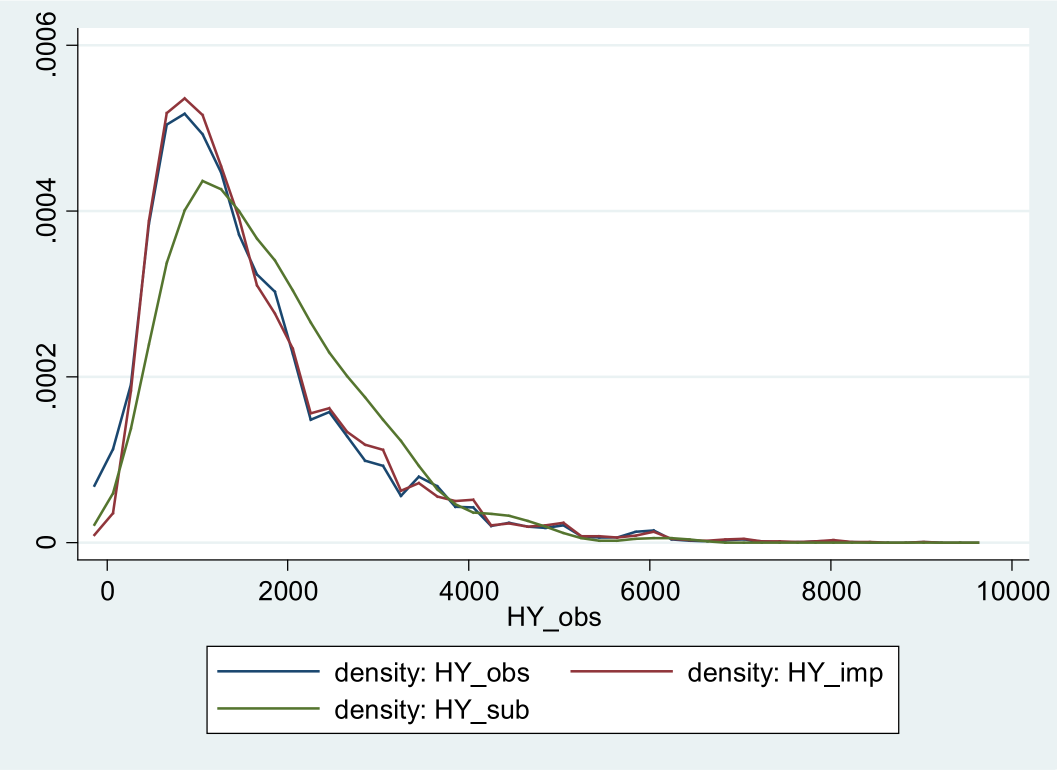


Table 7: Deciles of the distribution of different versions of household income (final estimates)

	dec1	dec2	dec3	dec4	dec5	dec6	dec7	dec8	dec9
HY_obs	541	700	900	1100	1300	1509	1840	2300	3000
HY_imp1	540	700	900	1100	1300	1500	1900	2400	3059
HY_imp2	532	700	900	1100	1300	1500	1900	2400	3100
HY_imp3	539	700	900	1100	1300	1500	1900	2400	3088
HY_imp4	540	700	900	1100	1300	1500	1900	2400	3059
HY_imp5	540	700	900	1100	1300	1500	1900	2397	3100
HY_sub1	671	895	1061	1353	1565	1827	2159	2541	3127
HY_sub2	482	728	967	1262	1522	1781	2124	2608	3227
HY_sub3	542	862	1118	1316	1572	1828	2110	2540	3085
HY_sub4	533	802	999	1214	1448	1770	2132	2457	3096
HY_sub5	527	808	1021	1258	1524	1758	2181	2457	3172

Looking at table 7, the graphical impression is confirmed. One can see that the distribution of the imputed household income, HY_imp1 to HY_imp5 is very close to the distribution of the observed income variable HY_obs. The largest differences can again be found for the right tail of the distribution, with the higher values of deciles 7 to 9 in the imputed data indicating that higher-income households are more inclined not to report their incomes.

This is buttressed by the distribution of imputed incomes for households without any observed information, HY_sub1 to HY_sub5. While there is now greater variance across imputation runs, the values of the deciles are larger than those for all households along the whole distribution.

5.7 Imputation of household income: a regression

As a check for the plausibility of the imputation, a regression model containing a few household variables is run once for observed (log) household income, using those households who reported exact information on net household income. The same model then is run on the five imputed data sets, using again Stata's mim command.

Table 8: Regression results for observed and imputed data, dependent variable log household income

	Observed Data			**Imputed Data**		
	Coeff.	Std.Err.	p-value	Coeff.	Std.Err.	p-value
BA-sample	-0.2864	0.0168	0.000	-0.3012	0.0151	0.000
household type (ref. 1-person-hh)						
couple, no kids	0.6791	0.0186	0.000	0.6731	0.0170	0.000
1-parent-household	0.6840	0.0188	0.000	0.6815	0.0180	0.000
couple, kids<16	0.9666	0.0204	0.000	0.9599	0.0189	0.000
couple, kids>16	0.9736	0.0264	0.000	0.9758	0.0231	0.000
couple, kids> and <16	1.1635	0.0334	0.000	1.1359	0.0302	0.000
more generation-hh	1.0156	0.0671	0.000	1.0038	0.0537	0.000
other	0.7885	0.0479	0.000	0.7724	0.0411	0.000
missing	1.0055	0.0517	0.000	0.9811	0.0411	0.000
housing type (ref. shared housing – WG)						
residential home	-0.0088	0.0946	0.926	-0.0039	0.0808	0.962
main tenant	0.0787	0.0280	0.005	0.1094	0.0244	0.000
subtenant, lodger	-0.1422	0.0425	0.001	-0.0692	0.0382	0.070
house owner	0.1946	0.0310	0.000	0.2101	0.0270	0.000
not paying rent	0.1216	0.0482	0.012	0.1855	0.0428	0.000
housing costs	0.0007	0.0000	0.000	0.0007	0.0000	0.000
square meters	0.0000	0.0000	0.000	0.0000	0.0000	0.000
index of deprivation	-0.0961	0.0037	0.000	-0.0990	0.0034	0.000
constant	6.6313	0.0359	0.000	6.6142	0.0330	0.000
obs	10232			12794		
R2	0.463					

The estimated coefficients look very similar. Thus, the completion of the data by imputation seems to have maintained the relationship between household income and the other variables included in the model. However, different from the individual earnings model, standard errors in the imputed data regression are throughout smaller in comparison to standard errors in the observed data regression. Thus, as would be expected, the completion of the household data has indeed led to an increase in the precision of the regression estimates, the influence of the uncertainty of the imputation procedure not being dominant here.

6 Concluding remarks

Multiple imputation is the preferred method to adjust for item nonresponse in surveys. The report summarizes the results of a research project implementing a multiple imputation approach for household income in the first wave of the German PASS data. The imputation is performed in several steps with the imputation of individual earnings serving as a prerequisite to the imputation of household income. To monitor the quality of the imputation, descriptive statistics and regression models for the completed data versions are discussed. The regression results are satisfactory in the sense that no obvious implausibities are detected and the imputations do not seem to distort the associations between central variables which are observed in the incomplete data.

Both on the individual level and on the household level, the imputed data for units without earnings or income information tended to be somewhat higher than observed earnings or income. This is in line with the finding of Frick et al. (2010), that “economically active household members are more common among PUNR (=partial unit non-response) and thus probably major contributions to overall household resources are understated” (p. 6).

Selective non-response within households could also be responsible for the gap observed between the sum of individual incomes of household members and total household incomes. This would call for a more refined model of individual earnings. Here, the PASS data offer a huge variety of potentially useful information. In addition to including more variables, labor supply could be modeled in a more sophisticated way, taking into account dependencies between household members. Of course, given the availability of longitudinal information in PASS, making use of the information contained in adjacent waves should be another step to deal with item nonresponse (see again Frick et al. 2010 for such an approach).

References

Beste, J., Eggs, J. and S. Gundert (2011), Instruments and Interview Programme. In: Bethmann, A., Gebhardt, D. (eds.), User Guide "Panel Study Labour Market and Social Security" (PASS). Wave 3. FDZ Datenreport 04/2011, Nürnberg.

Christoph, B., Müller, G., Gebhardt, D., Wenzig, C., Trappmann, M., Achatz, J., Tisch, A. and C. Gayer (2008), Codebuch und Dokumentation des "Panel Arbeitsmarkt und soziale Sicherung" (PASS). Band I: Einführung und Überblick. Welle 1 (2006/2007). FDZ Datenreport 05/2008, Nürnberg.

Drechsler, J. (2011), Multiple Imputation in Practice. A Case Study Using a Complex German Etablishment Survey. Advances in Statistical Analysis, Vol. 95, No. 1, S. 1-26.

Frick, J.R., Grabka, M.M. and O. Groh-Samberg (2010), Dealing with incomplete household panel data in inequality research. SOEPpapers 290, Berlin.

Hartmann, J., Brink, K., Jäckle, R. and N. Tschersich (2008): IAB-Haushaltspanel im Niedrigeinkommensbereich. Methoden- und Feldbericht. FDZ Methodenreport 07/2008, Nürnberg.

Ragunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and P. Solenberger (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology 27(1), 85-95.

Royston, P. (2005a), Multiple Imputation of Missing Values: Update. The Stata Journal, 5(2), 1-14.

Royston, P. (2005b), Multiple Imputation of Missing Values: Update of Ice. The Stata Journal, 5(4), 527-536.

Royston, P. (2007), Multiple Imputation of Missing Values: Further Update of Ice, With an Emphasis on Interval Censoring. The Stata Journal, 7(4), 445-464.

Royston, P. (2009), Multiple Imputation of Missing Values: Further Update of Ice, With an Emphasis on Categorical Variables. The Stata Journal, 9(3), 466-477.

Royston, P., Carlin, J.B. and I.R. White (2009), Multiple Imputation of Missing Values: New Features for mim. The Stata Journal, 9(2), 252-264.

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

Rubin, D.B. (1996), Multiple Imputation after 18+ Years. Journal of the American Statistical Association, 91(434), 473-489.

Seaman, S., White, I. (2008), Re-Analysis Using Inverse Probability Weighting and Multiple Imputation from Data from the Southampton Women's Survey, Cambridge, UK.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G. and A.J. Cohen (2006), Multiple Imputation of Missing Income Data in the National Health Interview Survey. Journal of the American Statistical Association, 101(475), Applications and Case Studies, 924-933.

Trappmann, M. (2011), PASS Background. In: Bethmann, A., Gebhardt, D. (eds.), User Guide "Panel Study Labour Market and Social Security" (PASS). Wave 3. FDZ Datenreport 04/2011, Nürnberg.

Trappmann, M., Müller, G., and A. Bethmann (2011), Design of the Study. In: Bethmann, A., Gebhardt, D. (eds.), User Guide "Panel Study Labour Market and Social Security" (PASS). Wave 3. FDZ Datenreport 04/2011, Nürnberg.

Oudshoorn, C.G.M., van Buuren, S. and van Rijckevorsel, J.L.A. (1999), Flexible multiple imputation by chained equations of the AVO-95 survey. Report PG/VGZ/99.045. Leiden, TNO Prevention and Health.

White, I.R., Rhian, D. and P. Royston (2010), Avoiding Bias Due to Perfect Prediction in Multiple Imputation of Incomplete Categorical Variables. Computational Statistics and Data Analysis 54, 2267-2275.

Yu, L.-M., Burton, A. and Rivero-Arias, O. (2007), Evaluation of Software for Multiple Imputation of Semi-Continuous Data. Statistical Methods in Medical Research, 16, 243-258.

Imprint

FDZ-Methodenreport 02/2012

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Iris Dieterich

Technical production
Iris Dieterich

Download
http://doku.iab.de/fdz/reporte/2012/MR_02-12_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:

Dr. Ursula Jaenichen,
Institute for Employment Research (IAB)
Regensburger Str. 104
D - 90478 Nuremberg
Phone: +49 (0)911-179-5415
Email: ursula.jaenichen@iab.de