DIW Berlin

German Institute
for Economic Research

Research Notes

29

Markus M. Grabka and
Joachim R. Frick

Imputation of Item-Non-Response on Income Questions in the SOEP 1984–2002

Berlin, October 2003

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

DIW Berlin

German Institute
for Economic Research

Königin-Luise-Str. 5
14195 Berlin,
Germany

Phone +49-30-897 89-0
Fax +49-30-897 89-200

www.diw.de

ISSN 1619-4551

The German Socio-Economic Panel Study

DIW Berlin
German Institute for Economic Research

DIW Research Note No. 29

IMPUTATION OF ITEM-NON-RESPONSE ON INCOME QUESTIONS IN THE SOEP 1984–2002

by

Markus M. Grabka and Joachim R. Frick *

Berlin, October 2003

Markus M. Grabka mgrabka@diw.de German Socio-Economic Panel Study (SOEP) DIW Berlin Koenigin-Luise-Strasse 5, D-14195 Berlin, Germany phone: +49-30-89789-339, Fax: +49-30-89789-109	Dr. Joachim R. Frick jfrick@diw.de German Socio-Economic Panel Study (SOEP) DIW Berlin Koenigin-Luise-Strasse 5, D-14195 Berlin, Germany phone: +49-30-89789-279, Fax: +49-30-89789-109

* We would like to thank Tasso Brandt for excellent research assistance.

Abstract:

This paper is an extension of a previous documentation by Butrica (1996) and describes the imputation methods for filling in item-non-response on income questions in the SOEP-based German part of the Cross-National-Equivalent File (CNEF) 1984-2002 which are identical to the $PEQUIV-files in the data distribution of the German Socio-Economic Panel Survey (SOEP). In contrast to cross-section surveys, the imputation of missing values in panel data like the SOEP can profit from longitudinal information which is available for the very same observation units from other points in time. The "row-and-column imputation procedure" developed by Little & Su (1989) considers longitudinal as well as cross-sectional information in the imputation process. This procedure is applied to the SOEP when deriving annual income variables, complemented by purely cross-sectional techniques.

1. Introduction

A common phenomenon in population surveys is the failure to collect complete information on interesting characteristics at individual and household level. In general, one can differentiate between unit and item-non-response. Unit-non-response results in the lack of any information on a given observation and turns out to be the strongest type of refusal – this issue is not being dealt with in this paper. If, however, only a subset of information is missing from an otherwise responding observation this failure is referred to as item-non response. The latter may be caused by a respondent's reservation to answer to a question that appears to be too sensitive (to him/her), or that affects confidentiality and privacy or simply from the fact that the correct answer is not known.

The German Socio-Economic Panel Study (SOEP)[1] distinguishes between three types of missing values: individuals may not know the answer to a question or may refuse to answer; the answer given may not make sense within the context of the question; or the answer given may later be found implausible and may be deleted by the interviewer. The codes for these item-non-response are:[2]

-1 or .A	no answer/do not know
-2 or .B	does not apply
-3 or .C	answer implausible and deleted subsequently.

By definition, any missing value of "*-2*" represents a valid answer and does not need to be corrected or imputed. Missing values coded as "*-1*" or "*-3*" are to be considered as "real" item-non-response in SOEP which are subject to imputation.

According to Rubin (1976) the mechanisms leading to missing data and item-non-response can be classified into three subgroups:

- Missing Completely at Random (MCAR),
- Missing at Random (MAR), and
- Missing Not at Random (MNAR).

MCAR means that the missing data mechanism is unrelated to the variables under study, whether missing or observed. The observed values are a random sample of the underlying population and any analysis on complete cases yields the very same results as the full data set

[1] For detailed description of the SOEP, see SOEP Group (2001).
[2] Missing Values in the SOEP are originally coded as -1, -2 and -3. Although the statistical software package SAS is able to deal with these codes, the standard missing values in SAS are coded as .A, .B and .C.

would have. An alternative and weaker version of the MCAR assumption is the Missing at Random (MAR) condition. The cause of the missing data is unrelated to the missing values, but may be related to the observed data. In other words MAR means that the missing values are related to either observed covariates or response variables. The third subgroup is Missing Not at Random (MNAR). MNAR occurs when the Missing mechanism depend on the actual value of the missing data. This is the most difficult condition to model for.

While MCAR and MAR may in principle be ignorable missing mechanism, MNAR requires adequate statistical treatment. There are in principle three ways how to deal with missing values:

(a) case-wise deletion,

(b) weighting and

(c) imputation.

A case-wise deletion may be either list-wise (complete cases only) or pair-wise. This very commonly used technique implies that cases are deleted which contain missing data in the variables which are relevant for the analysis being carried out. This procedure can substantially lower sample size, leading to a severe lack of power and may cause a bias in the analysis, depending on the underlying missing mechanism.

The second strategy to deal with missing values is weighting, i.e. "correcting" for an under-representation of certain characteristics by increasing the population weight of those observations which in fact did participate in the survey although they face a higher risk of non-response. A most relevant outcome of this strategy is an increase in the variance, producing less precise estimates.

The third strategy is imputation, where single and multiple imputation methods are being discussed. In contrast to single imputation techniques which provide only one estimate for each missing component, multiple imputation techniques return *m* complete datasets by imputing *m* times. It is argued that multiple imputation yields improved estimators compared to case-wise deletion, or single ad-hoc imputation methods, however, this technique may be shunned by less sophisticated users of micro-data.[3]

The advantage and also the aim of imputation is to complete a data set with "full" information for all observed individuals, which reduces bias in survey estimates and – from a point of a data user – also simplifies the analysis. Retaining all observations, independent from item-non-

[3] For an application of multiple imputation techniques to income data for Finland, see Spiess and Goebel (2003).

response, is supposed to yield an improved basis for (social) policy oriented analyses. However, it must be noted, that even a very sophisticated approach to substitute for non-response may not be sufficient to completely eliminate any bias resulting from it in the first place. As such, the adequate choice of the imputation technique is a problem by itself. Potential bias due to imputation may creep in due to "regression-to-the-mean effects" and a potential change in total variance – most likely a decline – can occur. This is of special relevance in surveys with a rather small number of observations or if small subgroups are affected above average.

In the following we present a detailed description of imputation methods as they are applied to substitute item-non-response on income questions in the SOEP (this documentation is based on the data distribution of 2003, ie. capturing SOEP data for the survey years 1984-2002).

2 Imputation procedures applied to the SOEP

2.1 The Row-and-Column imputation procedure

The imputation of item-non-response related missing income data in the SOEP follows a two step procedure. The general principle is to apply the so-called *row and column* imputation technique suggested by Little and Su (1989) (hereafter L & S) whenever longitudinal income data is available, and to run purely cross-sectional imputation techniques otherwise.

The row and column imputation procedure takes advantage of cross-sectional and longitudinal information of panel data by allowing information from other waves to be included and uses this information to compute a single value used to fill in missing values. In principle, the imputed value is the result of a combination of row (unit), column (period/trend)and a residual effect.

*imputation = (row effect) * (column effect) * (residual).*

The column effects are given by: $c_j = (19 * \bar{Y}_j) / \sum_{k=1}^{19} \bar{Y}_k$ and are calculated for each of the 19 waves of data (1984-2002) where j = 1, ..., 19 and $\bar{Y}_j$ is the sample mean income for year j.

The row effects are given by: $r_i = m_i^{-1} * \sum_{j=1}^{19} (Y_{ij} / c_j)$ and are computed for each sample member. Y_{ij} is the income for individual *i* in year *j* and m_i is the number of recorded months.

Sorting cases by r_i and matching the incomplete case *i* with information from the nearest complete case, say *l*, yields the imputed value:

$$i = [r_i] * [c_j] * [Y_{lj} / (r_l * c_j)] .$$

The three terms in brackets represent the *row*, *column*, and *residual* effects. The first two terms estimate the predicted mean, and the last term is the stochastic component of the imputation from the matched case.

Unlike traditional single imputation methods, the L & S imputation procedure incorporates residuals, based on sample data, into the imputed values. Another advantage that this procedure has over many other imputation methods is that it is rather easy to implement and does not require separate modeling for different patterns of missing data.

However, the empirical implementation of L & S in the case of SOEP fails in all those cases where a given income component is not observed in any other wave of data considered in the imputation process. This includes not only first time respondents, but also those observations for whom a given income variable has been surveyed for the very first time. In all those cases there is need for an alternative imputation procedure which is based on cross-sectional data only, i.e., on data observed from other units (individuals or household, respectively) in the very same wave.

2.2. Supplemental Cross-Sectional Imputation Methods

The supplemental imputation methods are based on cross-sectional information only. Depending on the relevance and complexity of the lacking income aggregate we apply various single imputation techniques:

- Logical Imputation: This imputation method is only feasible in cases where a straightforward link between a piece of missing information and at least one observed characteristic can be established. Here institutional or external information is used to impute missing amounts of those income components which are perfectly related to otherwise observed information, e.g., child benefit which is fixed per child, direct

housing support for owner occupiers which is related to the number of children and the construction year of the building, nursing care insurance which is fixed to the observed needs.

- Median Substitution takes place for income components which are of minor relevance in terms of the number of affected cases ($n < 10$) as well as with respect to the level (e.g. military service pay, maternity benefit). Median Substitution for Subgroups is performed for e.g. housing benefit for owner occupiers by household size. The choice for the median compared to the mean is to get a conservative estimate of the "true" value which is not biased by outliers as would be the case when relating to the population mean.
- Median Share Substitution is chosen if a link between two income variables can be established, e.g. the median share of the monthly labor earnings and the Christmas bonus in the private sector in Germany is about 35%. Resulting from this, any observation with a missing Christmas bonus in the private sector is assigned an imputed value given by the product of the individually observed labor income times the (median) share of 35%. This allows for more variation of the imputed income values than single median substitution would do and considers the individual income level as well.
- Regression-based substitution is used for more complex income constructs like e.g. "interest and dividends" or "individual labor income from first job"; in the latter case Mincer-type wage regressions are applied for imputation purposes.

See Appendix 1 for a complete overview of the various imputation techniques applied to the almost 50 SOEP income variables per year at individual and household level.

2.3. Quality of Imputation

Based on the income component "labor income from first job" we check the quality of the applied imputation techniques by comparing results for both, the L&S procedure and a purely cross-sectional imputation[4], with the actually observed information. Based on a random sample of approx. 1000 observations for which a positive value has been observed and who provide

[4] In this case a Mincer type regression is used (see section 3.1.1.)

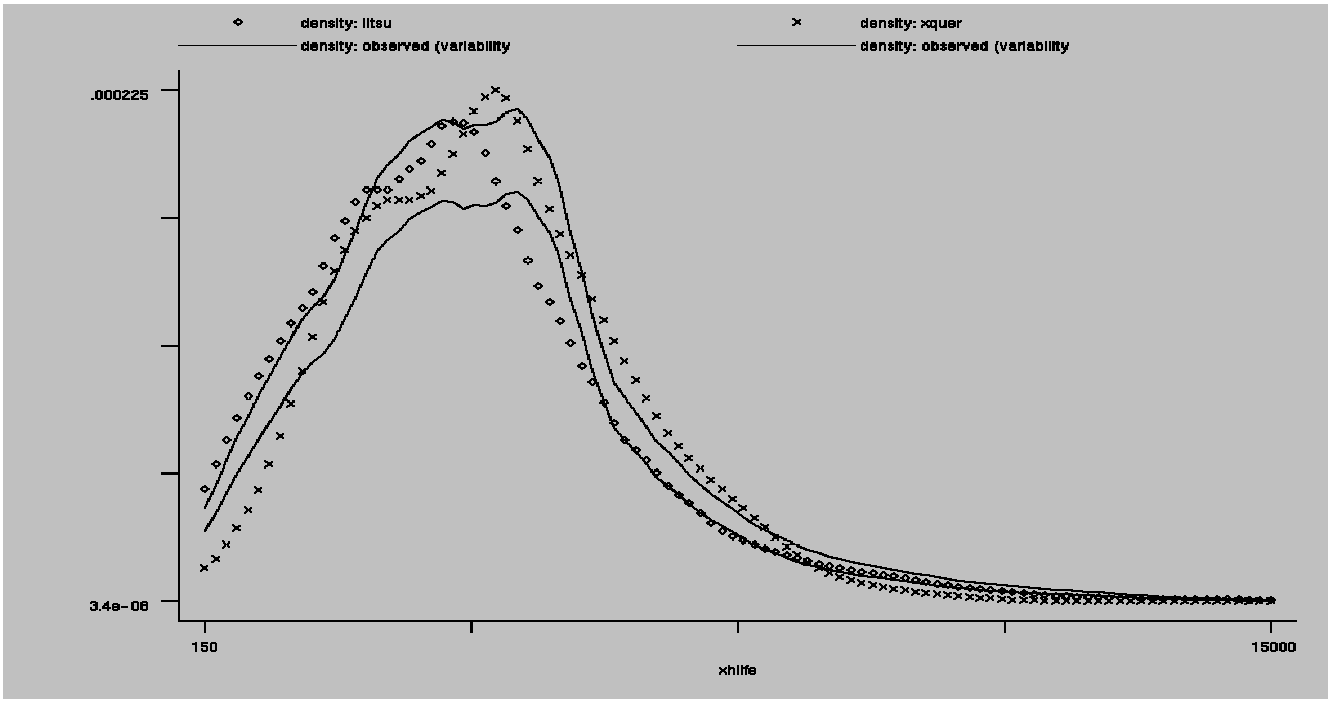
longitudinal information as a prerequisite for the L & S procedure, Figure 1 offers kernel density estimates for the three resulting distribution of "labor income from first job".

Mean and Median of the cross-sectional imputation procedure are in better compliance with those of the observed distribution than is the case for the L & S procedure. Although both techniques at first glance appear to yield rather reliable results, the distribution of the cross-sectional procedure is more "out-of-bounds" when comparing both imputation results to the upper and lower bounds of a 2-Sigma confidence band of the observed distribution which is due to the clearly understated variation. Another impressive illustration is incorporated in the significant understatement of the Gini coefficient: While the L & S procedure overstates inequality by about 9%, the cross-sectional approach understates the Gini by about 18%.

Concluding from this, one may argue that the L & S procedure, taking advantage from the individual's own record over time, yields more reliable imputation results than a purely cross-sectional approach does.[5] Although L & S is our preferred approach, there is obvious need for a purely cross-sectional imputation in case of lacking longitudinal data.

[5] This finding is in line with those of Spiess and Goebel (2003) based on survey and register data for Finland.

Figure 1: Kernel Density Estimates for "Individual Labor income from first job": Results from alternative imputation techniques vs. observed values



	Observed	Little & Su	X-Section
Mean	4 286	3 867	4 257
Median	4 000	3 501	4 180
Stdev	2 510	2 342	1 857
Dec. 90:10	5,14	5,81	3,63
Gini	0.3019	0.3284	0.2485
N	1,086	1,086	1,086

Note: Calculations are based on a random sample of 1086 observations (prgroup = 4) for which a positive value has been observed and who provide longitudinal information as a prerequisite for the Little & Su procedure.

Source: SOEP, Survey year 2001 (samples A-F), weighted results.

3. Detailed description of the supplemental cross-sectional imputation Methods

3.1 Individual Level Variables

3.1.1. Individual Labor Income from first job ($P2A03)

The imputation of individual labor income from first job ($P2A03) is based on 10 different regression models depending on the current employment status, the SOEP wave[6], and the region a person lives in (West- and East Germany, respectively):

Model 1, for current full-time and part-time employees, Wave A to G:

$P2A03 = [EDUCATION] + [NATIONALITY] + [INDUSTRY] + [COMPANY SIZE] + [AGE] + [AGE²] + [SEX] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT] + [PUBLIC/PRIVATE] + [JOB TENURE]

Model 2, for current full-time and part-time employees, Wave H to Q, in West Germany ($SAMPREG = 1):

$P2A03 = [EDUCATION] + [NATIONALITY] + [INDUSTRY] + [COMPANY SIZE] + [AGE] + [AGE²] + [SEX] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT] + [PUBLIC/PRIVATE] + [JOB TENURE]

Model 3, for current full-time and part-time employees, Wave H to Q, in East Germany ($SAMPREG = 2):

$P2A03 = [EDUCATION] + [NATIONALITY] + [INDUSTRY] + [COMPANY SIZE] + [AGE] + [AGE²] + [SEX] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT STATUS] + [PUBLIC/PRIVATE]

Model 4, for current full-time and part-time employees, Wave R:

$P2A03 = [EDUCATION] + [NATIONALITY] + [INDUSTRY] + [COMPANY SIZE] + [AGE] + [AGE²] + [SEX] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT STATUS] + [PUBLIC/PRIVATE] + [REGION]

Model 5, for current full-time and part-time employees, Wave S

$P2A03 = [EDUCATION] + [NATIONALITY] + [INDUSTRY] + [COMPANY SIZE] + [AGE] + [AGE²] + [SEX] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT STATUS] + [PUBLIC/PRIVATE] + [REGION] + [SAMPLE]

Model 6, other than current full-time or part-time employees, Wave A through G:

$P2A03 = [EDUCATION] + [NATIONALITY] + [AGE] + [AGE²] + [SEX] + [LABOR FORCE] + [MARITAL STATUS] + [#CHILDREN]

[6] Wave A (=survey year 1984) thru Wave S (=survey year 2002); the East German Sample starts with wave G (=survey year 1990).

Model 7, other than current full-time or part-time employees, Wave H through Q, $SAMPREG = 1:

$P2A03 = [EDUCATION] + [NATIONALITY] + [AGE] + [AGE²] + [SEX] + [LABOR FORCE] + [MARITAL STATUS] + [#CHILDREN]

Model 8, other than current full-time or part-time employees, Wave H through Q, $SAMPREG = 2:

$P2A03 = [EDUCATION] + [NATIONALITY] + [AGE] + [AGE²] + [SEX] + [LABOR FORCE] + [MARITAL STATUS] + [#CHILDREN]

Model 9, other than current full-time or part-time employees, Wave R:

$P2A03 = [EDUCATION] + [NATIONALITY] + [INDUSTRY] + [COMPANY SIZE] + [AGE] + [AGE²] + [SEX] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT STATUS] + [PUBLIC/PRIVATE] + [REGION]

Model 10, other than current full-time or part-time employees, Wave S

$P2A03 = [EDUCATION] + [NATIONALITY] + [INDUSTRY] + [COMPANY SIZE] + [AGE] + [AGE²] + [SEX] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT STATUS] + [PUBLIC/PRIVATE] + [REGION] + [SAMPLE]

Covariates:

[NATIONALITY]: dummy variable for nationality: (1) = German, (0) otherwise

[INDUSTRY]: set of 4 dummy variables for the current industry code ($NACE)
(1) primary sector;
(2) manufacturing;
(3) service sector;
(4) financial services

[COMPANY SIZE]: size of company where respondent is employed (BETR$$)

[SEX]: dummy variable for gender: (0) = female, (1) = male

[AGE]: age of respondent (derived from GEBJAHR)

[AGE²]: AGE*AGE

[EDUCATION]: 3 dummy variables based on number of years in education ($BILZEIT): (1) le 11 years; (2) 12-14 years; and (3) more than 14 years in education

[OCCUPATION]: set of 11 dummy variables (based on RP4001 to RP4005)
(1) untrained worker, semi-trained worker, untrained employee;
(2) trained worker, apprentice/volunteer, low-level civil service;
(3) foreman, (semi-)trained employee, (qualified) professional, middle-level civil service;
(4) master craftsman, industry foreman, (highly qualified) professional, high-level civil service;
(5) managerial, executive civil service;
(6) self-employed farmer;
(7) free-lance professional;
(8) self-employed without other coworkers;

(9) self-employed with le 9 coworkers;
(10) self-employed with more than 9 coworkers;
(11) family member working for relative

[MARITAL STATUS]: dummy for marital status ($FAMSTD):
(1) if married/married, but separated, (0) otherwise

[EMPLOYMENT]: set of dummy variables for current employment status (e.g.RP12)
(1) full-time employed, vocational training;
(2) part-time employed, marginal part-time;
(3) otherwise

[PUBLIC/PRIVATE]: dummy variable for public and private sector (OEFFD$$):
(1) public sector, (0) private sector

[JOB TENURE]: job tenure of respondent ($ERWZEIT)

[LABOR FORCE]: set of dummy variables based on the current labor force status (LFS$$):
(1) non-working (nw), nw – age 65 and older, nw – in education-training, nw – military/civil service;
(2) nw – maternity leave;
(3) nw – unemployed;
(4) nw – but sometimes sec. Job;
(5) nw – but past 7 days;
(6) nw – but reg. sec. job

[#CHILDREN]: number of children in household ($KZAHL)

[REGION]: dummy variable to differentiate East and West German Households ($SAMPREG)

[SAMPLE]: set of 7 dummy variables indicating the SOEP-sub-sample A to G.

3.1.2. Income from Self-employment ($P2B03)

The imputation of income from self-employment is based on 4 different regression models.

Model 1 wave A to H:

$P2B03 = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT]

Model 2 wave I to R:

$P2B03 = [SEX] + [AGE] + [AGE²] +[EDUCATION] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT] + [REGION]

Model 3 since wave S:

$P2B03 = [SEX] + [AGE] + [AGE²] +[EDUCATION] + [OCCUPATION] + [MARITAL STATUS] + [EMPLOYMENT] + [REGION] + [SAMPLE]

Model 4 for cases which had negative imputes resulting from the standard-models, a simplified regression model was applied:

wave A to H: $P2B03 = [AGE] + [AGE²] + [SEX]
since wave I: $P2B03 = [AGE] + [AGE²] + [SEX] + [REGION]

Covariates: (see income from first job)

3.1.3. Income Second Job ($P2C03)

Imputation based on a:
- Median-substitution for wave A to G
- Median-substitution by $SAMPREG since wave H.

3.1.4. Income from own pensions (e.g. rp7901)

The imputation of income from own pensions is based on 3 different regression models.

Model 1 wave A to H:
[OWN PENSION] = [SEX] + [AGE] + [AGE²] + [EDUCATION]
Model 2 wave I to R separately for West and East Germany :
[OWN PENSION] = [SEX] + [AGE] + [AGE²] + [EDUCATION]
Model 3 since wave S :
[OWN PENSION] = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [SAMPLE]

Covariates: (see income from first job)

3.1.5. Income from widow or orphan pensions (e.g. rp7910)

The imputation of income from own pensions is based on 3 different regression models.

Model 1 wave A to H:
[WIDOW PENSION] = [SEX] + [AGE] + [AGE²] + [EDUCATION]
Model 2 wave I to R separately for West and East Germany :
[WIDOW PENSION] = [SEX] + [AGE] + [AGE²] + [EDUCATION]
Model 3 since wave S :
[WIDOW PENSION] = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [SAMPLE]

Covariates: (see income from first job)

3.1.6. Company Pensions ($P2P03)

Imputation based on a regression model, since 2002 separately for West and East Germany
$P2P03 = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [SAMPLE]

Covariates: (see income from first job)

3.1.7. Private Pensions ($P2Q03)

Imputation based on a regression model, since 2002 separately for West and East Germany
$P2Q03 = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [SAMPLE]

Covariates: (see income from first job)

3.1.8. Unemployment Benefit ($P2F03) and Subsistence Allowance ($P2H03)

The imputation of unemployment benefits make use of two different imputation techniques depending on the availability of income from first job ($P2A03) which is the basis for calculating this transfer. The same strategy is used for subsistence allowances due to comparable regulations.

1. Imputation based on a Median-share substitution given by

Median[($P2F03) / $P2A03] * FIRSTJ$,

where *$FIRSTJ* is the individual income aggregate representing income from first job consisting both of original and imputed values of $P2A03.

2. Imputation based on a regression model for cases where $FIRSTJ is missing

Wave A to R: $P2F03 = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [REGION]
Wave S: $P2F03 = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [REGION] + [SAMPLE]

Covariates: (see income from first job)

3.1.9. Unemployment Assistance ($P2G03)

The imputation of unemployment assistance make use of three different imputation techniques depending on the availability of income from first job ($P2A03).

1. Imputation based on a Median-share substitution given by

Median[($P2G03) / $P2A03] * FIRSTJ$,

where *$FIRSTJ* is the individual income aggregate representing income from first job consisting both of original and imputed values of $P2A03.

2. If *$FIRSTJ* is not applicable the imputation is based on a Median-share substitution given by

Median[($P2G03) / $P2A03] * ARBGELD$,

where ARBGELD$ is an aggregated variable representing 'unemployment benefit' and consisting both of (1) original and (2) imputed values of $P2F03.

3. If *$FIRSTJ* and ARBGELD$ are not applicable the imputation is based on a regression model:

Wave A to R: $P2G03 = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [REGION]
Wave S: $P2G03 = [SEX] + [AGE] + [AGE²] + [EDUCATION] + [REGION] + [SAMPLE]

Covariates: (see income from first job)

3.1.10. Old Age Transition Payment ($P2I03)

Imputation based on a Median-substitution (due to very low number of observations with item-non-response).

3.1.11. Maternity/Paternity Assistance ($P2J03)

Imputation based on a Median-substitution (due to very low number of observations with item-non-response).

3.1.12. Student Assistance ($P2K03)

Imputation based on a Median-substitution for students who

- are living together with their parents ($STELL >= 3) or
- are living apart from their parental household (approximated by $STELL = 0, 1, 2).

3.1.13. Military/ Civilian Service Pay ($P2L03)

Imputation based on a Median-substitution (due to very low number of observations with item-non-response).

3.1.14. Private Transfers from persons outside the household ($P2M03)

Imputation based on a Median-substitution (due to very low number of observations with item-non-response).

3.1.15. Alimony ($P2O03)

Imputation based on a Median-substitution (due to very low number of observations with item-non-response).

3.1.16. 13th monthly salary (e.g. BP5902)

Imputation based on a Median-share substitution given by

$$Median\ (\$x\ /\ \$P2A03)\ *\ \$FIRSTJ_i,$$

where *$x* is the 13th monthly salary, *$P2A03* is the income from first job for the observed cases, and *$FIRSTJi* is the individual income aggregate representing income from first job. The latter consists of both, original and imputed values of *$P2A03* for cases where the 13th monthly salary is missing.

3.1.17. 14th monthly salary (e.g. BP5904)

Imputation based on a Median-share substitution given by

$$Median (\$x / \$P2A03) * \$FIRSTJ_i,$$

where *$x* is the 14th monthly salary, *$P2A03* is the income from first job for the observed cases, and *$FIRSTJi* is the individual income aggregate representing income from first job. The latter consists of both, original and imputed values of *$P2A03* for cases where the 14th monthly salary is missing.

3.1.18. Christmas Bonus (e.g. BP5606)

Imputation based on a Median-share substitution given by

$$Median (\$x_p / \$P2A03_p) * \$FIRSTJ_{ip},$$

differentiated for those currently employed in the public sector (OEFFD$$ = 1) or private sector (OEFFD$$ = 0), where $\$x_p$ is the Christmas bonus and $\$P2A03_p$ the income from first job for the observed cases separately for currently employed in the public or private sector, and $\$FIRSTJ_{ip}$ is the individual income aggregate representing income from first job consisting both of original and imputed values of $\$P2A03_p$ for cases where the Christmas bonus is missing.

3.1.19. Vacation Bonus (e.g. BP5908)

Imputation based on a Median-share substitution given by

$$Median (\$x_p / \$P2A03_p) * \$FIRSTJ_{ip},$$

differentiated between currently employed in the public sector (OEFFD$$ = 1) or private sector (OEFFD$$ = 0), where $\$x_p$ is the vacation bonus and $\$P2A03_p$ the income from first job for the observed cases separately for currently employed in the public or private sector, and $\$FIRSTJ_{ip}$ is the individual income aggregate representing income from first job consisting both of original and imputed values of $\$P2A03_p$ for cases where the vacation bonus is missing.

3.1.20. Profit Sharing (e.g. BP5910)

Imputation based on a Median-share substitution given by

$$Median (\$x / \$P2A03) * \$FIRSTJ_i,$$

where *$x* is the profit sharing and *$P2A03* the income from first job for the observed cases, and $\$FIRSTJ_i$ is the individual income aggregate representing income from first job consisting both of original and imputed values of *$P2A03* for cases where profit sharing is missing.

3.1.21. Other Perks (e.g. BP5912)

Imputation based on a Median-share substitution given by

$$Median (\$x / \$P2A03) * \$FIRSTJ_i,$$

where *$x* is other perks and *$P2A03* the income from first job for the observed cases, and $\$FIRSTJ_i$ is the individual income aggregate representing income from first job consisting both of original and imputed values of *$P2A03* for cases where other perks is missing.

3.2. Household Level Variables

3.2.1. Child Benefit (e.g. RH4803)

Imputation is based on a logical substitution, where legally fixed amounts, varying with the number of children in the household ($KZAHL) are assigned.

Example (wave R):

```
if rh4803 = -1,-3 then do;
        if rkzahl   = 1 then rh4803= 270;
   else if rkzahl   = 2 then rh4803= 270 + 270;
   else if rkzahl   = 3 then rh4803= 270 + 270 + 300;
   else if rkzahl  gt 3 then rh4803= 270 + 270 + 300 + ((rkzahl -3) * 350);
end;
```

3.2.2. Nursing Care Insurance/Amount Of Permanent Disability Support (e.g. RH4609)

Imputation is based on a logical substitution, where a legally fixed amount is assigned. Due to the lack of further information a medium degree of permanent disability (out of a maximum of three so-called “Pflegestufen”) is assumed.

Example (wave R):

```
if rh4609 = -1,-3 then rh4609 = 800;
```

3.2.3. Social Assistance (HLU e.g. RH4703) and special circumstances benefit (HBL e.g. RH4706)

The imputation of Social Assistance and special circumstances benefit make use of two different imputation techniques. This is due to changes in the underlying questionnaire.

1. Wave A to H and since wave R (e.g. rh4703 and rh4706)

Imputation based on a Median-substitution by subgroups given by the household size ($HHGR).
Example (for wave R):

```
if rh4703 = -1,-3 then do;
    if rhhgr =   1 then rh4703 = (median(rh4703)) where rhhgr = 1;
    if rhhgr =   2 then rh4703 = (median(rh4703)) where rhhgr = 2;
    if rhhgr =   3 then rh4703 = (median(rh4703)) where rhhgr = 3;
    if rhhgr =   4 then rh4703 = (median(rh4703)) where rhhgr = 4;
    if rhhgr ge 5 then rh4703 = (median(rh4703)) where rhhgr ge 5;
end;
```

2. Wave I to Q

The exact amount of money received according to these two types of social benefits has not been asked in the SOEP for the years 1992 to 2000. The only information reported was whether a households received these incomes and eventually the number of months with receipt.

2a) Wave I to K:
The basic statutes governing receipt and level of social welfare benefits were incorporated into an estimation routine. The first step in estimating these benefits was to identify individuals within households whose income levels would qualify them to receive social welfare benefits. A “needy” index was created to identify these individuals. Persons in households who earned four times the amount specified in the basic statues were assigned a “needy” value of zero. The household needs were determined by the sum of the needs of all household members with additional supplements for the elderly, single parent households, and those gainfully employed. The estimation routine used is based on the work of Andreß et al. (1995). The estimation routine produces estimates for social assistance (HLU) and special circumstances assistance (HBL). If a household received both HBL and HLU benefits then the HBL benefits were assumed to be one-time payments to the household and were assigned and amount equal to 10% of the HLU benefits. Social welfare benefits for the current survey year were imputed only for households that reported receiving social welfare benefits (HLU or HBL) in the previous survey year. The imputed value for social welfare benefits is the sum of the HBL and the HLU benefits (see Butrica 1996).

2b) Wave L to Q:

Imputation is based on median substitution for three subgroups according to receipt of HLU, or HBL or both in a given wave.

3.2.4. Income from Rental and Leasing (e.g. RH41) and Operation/Maintenance Costs (e.g. RH4201)

Imputation based on Median-substitution by three subgroups (low, medium and high monthly household net income).
Example (for wave R):

```
if rh41 = -1,-3 then do;
   if rh49 lt 2500                then rh41 = (median(rh41)) where rh49 lt 2500;
   if rh49 >2500-5000< then rh41 = (median(rh41)) where rh49 >2500-5000<;
   if rh49 ge 5000                then rh41 = (median(rh41)) where rh49 ge 5000;
end;
```

3.2.5. Housing Benefit (e.g. RH4606)

Housing benefit is treated separately for owner occupiers and tenants.

1. Owner occupiers: Imputation based on a Median-substitution
2. Tenants: Imputation based on a regression model

   Housing benefit = [HOUSEHOLD SIZE] + [SIZE OF HOUSING UNIT] + [MONTHLY INCOME] + [AMOUNT OF RENT]

Covariates:

[HOUSEHOLD SIZE]:	Number of persons living in household ($HHGR)
[SIZE OF HOUSING UNIT]:	size of housing unit in squared meters ($WOHNFL)
[MONTHLY INCOME]:	current monthly net household income (e.g. RH49)
[AMOUNT OF RENT]:	amount of rent minus heating costs ($MIETEG)

3.2.6. Income from interest and dividends(e.g. RH4401)

Imputation based on four different regression models:

Model 1, Wave A to H:

LOG([INTEREST INCOME]) = [OWNERSHIP] + [SELFEMPLOYED] + [INCOME] + [EDUCATION] + [AGE] + [#CHILDREN] + [SEX] + [CAPITAL ASSETS]

Model 2, Wave I to R, for West-Germany only ($SAMPREG = 1):

LOG([INTEREST INCOME]) = [OWNERSHIP] + [SELFEMPLOYED] + [INCOME] + [EDUCATION] + [AGE] + [#CHILDREN] + [SEX] + [CAPITAL ASSETS]

Model 3, Wave I to R, for East-Germany only ($SAMPREG = 2):

LOG([INTEREST INCOME]) = [OWNERSHIP] + [SELFEMPLOYED] + [INCOME] + [EDUCATION] + [AGE] + [#CHILDREN] + [SEX] + [CAPITAL ASSETS]

Model 4, Wave S:

LOG([INTEREST INCOME]) = [OWNERSHIP] + [SELFEMPLOYED] + [INCOME] + [EDUCATION] + [AGE] + [#CHILDREN] + [SEX] + [CAPITAL ASSETS] + [REGION] + [SAMPLE]

Independent Variable:

[INTEREST INCOME]: An aggregated metric variable for income from interest and dividends based on information taken from a metric variable (e.g., RH4401) and a categorical variable (e.g., RH4402). The latter was transformed into a metric variable assuming a random uniform distribution within each of the categories.

Covariates:

[OWNERSHIP]:	dummy variable: (1) if owner occupier, (0) otherwise ($EIGEN)
[SELFEMPLOYED]:	dummy variable based on information on the household head: (1) if e.g. RP4002 >= 1, (0) otherwise
[INCOME]:	7 different dummy variables for current monthly household net income (e.g. RH49) differentiation between categories (1) 1-1000 DM (2) 1001-2000 DM; (3) 2001-3000 DM; (4) 3001-5000 DM; (5) 5001 -7500 DM; (6) > 7500DM; additionally, an indicator variable for cases with missing values (.A or .C) on RH49
[EDUCATION]:	3 dummy variables based on number of years in education (head) ($BILZEIT): (1) le 11 years; (2) 12-14 years; and (3) more than 14 years in education
[AGE]:	age of respondent (derived from GEBJAHR)
[#CHILDREN]:	number of children in household ($KZAHL)
[SEX]:	dummy variable for gender: (0) = female, (1) = male
[CAPITAL ASSETS]:	set of 5 dummy variables: e.g. QH4301 = savings account, QH4302 = building society savings, QH4303 = life insurance, QH4304 = securities, bonds and stocks, QH4305 operating assets.
[REGION]:	dummy variable to differentiate East and West German Households ($SAMPREG)
[SAMPLE]:	set of 7 dummy variables indicating the sample affiliation for sample A to G.

3.2.7. Direct Housing support for owner occupiers (e.g. RH3904)

Imputation is based on a logical substitution, where legally fixed amounts, varying with number of children in household ($KZAHL) and the year of construction ($BAUJ) are assigned.
Example (wave S):

```
if sh3904 = -1,-3 then do;
if sbauj in (1,2,3,4,5) then do;
   if skzahl =  .  then sh3904 = 2500;
   if skzahl ge 0 then sh3904 = 2500 + (1500 * skzahl);
  end;
  if sbauj ge 6 then do;
   if skzahl =  .  then sh3904 = 5000;
   if skzahl ge 0 then sh3904 = 5000 + (1500 * skzahl);
  end;
 end;
```

3.2.8. Taxes and Social security contributions

Taxes and social security contributions are fully simulated in the SOEP. For detailed description see Schwarze (1995).

3.2.9. Imputed rental value

Imputed rental value of owner occupied housing is a fully simulated information. A fictitious market rent for owner occupiers is estimated by using a hedonic regression model of gross rent in terms of square meters (not including heating) actually paid by main tenants in privately financed housing (without social housing and households with reduced rent). Independent variables include indicators describing the condition of the house, the year of construction, size of dwelling, length of occupancy, community size and disposable income. Applying these regression coefficients to the population of owner occupiers yields an estimate of the *gross* value at market prices (without costs for heating and warm water). Further deducting owner-specific costs for taxation, maintenance and operating costs as well as interest on mortgages yields a *net* value which can be interpreted as the appropriate income advantage of owner-occupied housing. In case of owner related costs exceeding the income advantage (especially at the beginning of the mortgage repayment period), IR is assigned a value of zero.[7]

3.2.10. Windfall income (e.g. RH4505)

Imputation based on a median substitution, only. Given the extreme volatility of this income component there is *no* application of the L & S procedure even in case of existing longitudinal data.

[7] For a more detailed description and an empirical application of these imputed rental values see Frick and Grabka (2003).

References

Andreß , H.-J., G. Lipsmeier, R. Samson, W. Strengmann-Kuhn. (1995): Income Analysis with Data from the Socio-Economic Panel. Working Paper No. 22 of the DFG-Project "Provision Strategies of Lower Income Private Households." Department of Sociology, University of Bielefeld.

Butrica, B.A. (1996). Imputation Methods for Filling in Missing Values in the PSID-GSOEP Equivalent File 1980-1994. Working Paper.

Frick, Joachim R. and Grabka, Markus M. (2003): Imputed Rent and Income Inequality: A Decomposition Analysis for the U.K., West Germany, and the USA. Review of Income and Wealth. forthcoming December 2003.

Little, R. J. A. & Su, H.-L. (1989). Item Non-Response in Panel Surveys. In D. Kasprzyk, G. Duncan and M. P. Singh (Eds.), Panel Surveys. New York: John Wiley.

Rubin, D.B. (1976). Inference with missing data. Biometrika, 63, 581-592.

SOEP Group (2001): "The German Socio-Economic Panel (SOEP) after more than 15 years – Overview," Vierteljahrshefte zur Wirtschaftsforschung, 70(1), 7-14.

Spieß, M. and J. Goebel (2003): A comparision of different imputation strategies with respect to income related questions. Paper presented on the Chintex final conference "Harmonisation of Surveys and Data Quality", Wiesbaden, 26 and 27 May 2003.

Schwarze, J. (1995). "Simulating German Income and Social Security Tax Payments Using the GSOEP." Cross-National Studies in Aging. Program Project Paper No. 19, Center for Policy Research, The Maxwell School. Syracuse, NY: Syracuse University.

APPENDIX 1:

Aggregated SOEP annual income variables (in CNEF-SOEP and $PEQUIV), the underlying original survey information from SOEP income components, and the respective type of imputation in case of missing values due to item-non-response

Income Aggregate	**Input (=original survey information with aggregation level)**	**primary imputation technique[1)]**	**secondary imputation technique**
I11101$$ (Household Pre-Government Income)	sum (I11103$$ + I11104$$ + I11106$$ + I11117$$)	see respective input variables below	see respective input variables below
I11103$$ (Household labour income) [10 inputs]	• Aggregated Household Labour Income		
	• first job	L & S	R
	• second job	L & S	M-G
	• self-employment	L & S	R
	• 13th monthly payments	L & S	M-S
	• 14th monthly payments	L & S	M-S
	• Christmas bonuses	L & S	M-S
	• vacation/holiday pay	L & S	M-S
	• profit sharing, premiums	L & S	M-S
	• other bonuses	L & S	M-S
	• military service pay	L & S	M-S
		L & S	M
I11104$$ (Household asset income) [3 inputs]	• income from rent and lease	L & S	M-G
	• minus operating & maintenance costs	L & S	M-G
	• interest & dividends	L & S	R
I11106$$ (Household private transfers) [2 inputs]	• Aggregated Household private transfers		
	• individual private transfers	L & S	M
	• alimony	L & S	M
I11117$$ (Household private pensions) [6 inputs]	• Aggregated Household private pensions		
	• own company retirement plan	L & S	R
	• own pension for public employees	L & S	R
	• other own pension	L & S	R
	• widow company retirement plan	L & S	R
	• widow pension for public employees	L & S	R
	• other widow pension	L & S	R

1) L&S = Little & Su 1989
M = Median Substitution
M-G = Median Substitution by Subgroups
M-S = Median share Substitution
R = Regression based imputation
Fixed = Fixed amounts

Income Aggregate	**Input (=original survey information with aggregation level)**	**primary imputation technique[1)]**	**secondary imputation technique**
I11102$$ (Household Post-Government Income)	= sum (I11101$$ + I11107$$ + I11108$$ - I11109$$)	see respective input variables below	see respective input variables below
I11107$$ (Household total public transfers) [6 inputs] [6 inputs]	• Aggregated Household private pensions Individual level: • higher education grants • maternity benefits • unemployment benefit • unemployment assistance • subsistence allowances • early retirement benefit Household level: • housing benefit for renters • housing benefits for owner-occupiers • child benefit • social assistance • special help income • nursing care insurance • direct housing support for owners	 L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S	 M-G M M-S M-S M-S M-S R M-G Fixed M-G M-G Fixed Fixed
I11108$$ (Household Social Security Pensions) [12 inputs]	• Aggregated Household social security Pensions • own GRV pension • own minors pension • own civil servants pension • own war victims pension • own farmers pension • own accident insurance pension • widow GRV pension • widow minors pension • widow civil servants pension • widow war victims pension • widow farmers pension • widow accident insurance pension	 L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S	 R R R R R R R R R R R R R
I11109$$ (Household federal Taxes and SSC)	complete imputation based on a micro-simulation programme	completely simulated	
I11105$$ (Imputed rental value)	complete imputation based on a hedonic regression estimation	completely simulated	
I11118$$ (Windfall income)	revenues from inheritances, lotteries, etc. (> 5000 DM / 2.500 EUR)	M	