

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (ii): Generalized software packages for statistical data editing, their evaluation

MULTIPLE IMPUTATIONS WITH SOLAS VERSION 1.1

Submitted by Statistics Denmark¹

Contributed paper

I. INTRODUCTION

1. Statistics Denmark uses a lot of resources on data editing. In this paper, the product SOLAS is tested on real data to evaluate how well it performs. The data are from the Danish Labour Force Survey (LFS) 1996, 2nd quarter. Data are based on a sample from which e.g. unemployment is estimated. Here we ignore the post stratification process but concentrate on making imputations on the missing data in the sample. We know which persons in the sample were interviewed and which persons were not interviewed. From the registers we have a lot of information about all respondents e.g. age, education and marital status. We try to impute "Income before tax" for the part of the sample that we did not interview. "Income before tax" is a register variable, which is known for the entire sample, but imputing a known variable enables us to evaluate the imputation process.

II. SOLAS FOR MISSING DATA ANALYSIS

2. SOLAS is a computer program produced by Statistical Solutions Ltd. in Ireland for imputation. We have tested SOLAS version 1.1. Besides the imputation tools, the program also contains standard statistical tools. The interface is a nice Windows interface very much like SAS/INSIGHT. The only tools in SOLAS not available in SAS/INSIGHT are the imputations routines. For a complete list of the standard statistical tools, see the SOLAS User Reference guide.

3. SOLAS was originally designed for imputation in biostatistical research. It supports both continuous and ordinal data for imputations but the imputation functions are more used on continuous data. It can also handle both single-observation data as well as longitudinal data. SOLAS imputation tools:

- **Group Mean Imputation.** This is the basic mean imputation where each missing value is exchanged with the group mean. Only one variable can be selected as a grouping variable and the user has to select the variable.
- **Last value carried forward.** SOLAS can impute longitudinal dataset by replacing missing value with a previous value.
- **Random Hot Deck.** Missing values are imputed by SOLAS by values from respondents who are similar to the non-respondent with regard to some auxiliary variables. The user must specify which auxiliary variables to use and the order of importance. If no complete match can be found the least important variable is dropped, if there still is no complete match the 2nd least important auxiliary variable is dropped and so on until an exact match is found or there is no more auxiliary variable to drop. The user

¹ Prepared by Bjorn Steen Larsen and Birger Madsen.

can select if the first exact match should be used or a random exact match should be used. It could have been nice to have a more flexible matching procedure i.e. finding the respondent which matches with most variables.

- **Multiple Imputation.** This is the most sophisticated imputation method in SOLAS. Based on the procedure developed by Rubin in the 1970s. Basically, SOLAS impute several values for each missing value in the dataset. Estimations can be made on each set of estimated values. The method helps to identify the estimation error, which is the result of the imputed values. The method is described as "SOLAS applies an implicit model approach (using a logistic regression model) based on propensity scores and an approximate Bayesian bootstrap to generate the imputations. The multiple imputations are independent repetitions from a posterior predictive distribution for the missing data given the observed data". Broadly speaking SOLAS estimates a non-response probability from the auxiliary variables. Then it makes five groups based on the 20%, 40%, 60%, 80% quantiles of the estimated probabilities. All missing value in each group are filled with values from random respondents in the same group.
4. Except for Multiple Imputations all SOLAS imputations algorithms can be implemented in a few lines of SAS code. The user interface of SOLAS which is standard Windows works good, but compared to SAS/INSIGHT it feels awkward. The driving force behind SOLAS is the renown statistician Dr. Rubin who initially developed the multiple imputation method.

III. DATA

5. The data in this report come from the Danish Labour Force Survey (LFS), 1996 2nd quarter. The survey is conducted by drawing a random sample of 10600 15-67 year old persons who are not registered as unemployed and 5000 registered unemployed 15-67 year old persons. The survey is a rotating panel survey where the selected persons are interviewed in three different quarters by a special rotation pattern. The rotating panel structure will be ignored here, because only data from one quarter are used. The selected persons are interviewed by phone if possible, otherwise they are reached by mail questionnaires.

6. The design of the survey affects the estimation of parameters but for our imputation problem it is of little importance so we will not go into further detail about this. Of the persons in the original sample, 10 have left the domain of study because they either had died or emigrated. We have also removed 8 outliers, all people with a before tax income of more than 1,000,000 DKK. The sample we have used consists therefore of 15,542 people. Among these, 11,389 (73%) were successfully interviewed and 4,153 (27%) were not successfully interviewed. During the interviews, Statistics Denmark received a lot of information about the interviewed respondents, but we only have register information about the non-interviewed. One register variable is "Income before taxes". In this evaluation we presume that "Income before taxes" is not a register variable and we try to impute values for the 4,153 non-interviewed based on the successfully interviewed respondents and the available register information.

7. The following standard register information were also available (besides "Income before tax"):

- i) Sex,
- ii) Age,
- iii) Living alone,
- iv) Marriage status,
- v) Children under 18 living at home (yes/no),
- vi) Area of living (12 classes).
- vii) Education, (non, short, long or missing information)
- viii) Type of job (government, manufactory, private sector, other i.e. unemployed).
- ix) Registered as unemployed.

8. We have made a logistic regression of response and a linear regressing of income to be able to find out the variables that are best for grouping (Random Hot Deck) or best possible covariates (Multiple Imputation).

IV. THE STRUCTURE OF NON-RESPONSE

9. We have used a logistic regression model to find the variables, which can explain most of the response rates. The following factors were very significant. The most significant are placed at the top.

Type 3 Wald tests from the logistic regression

Source	DF	Chi-Sq.	Pr>Chi-Sq
Level of Education	3	405,6	0,0001
Sector of Employment	3	187,5	0,0001
Living alone status	1	232,2	0,0001
Employment status (register)	1	91,4	0,0001
Area of living	11	35,1	0,0001
Marriage status	1	19,2	0,0001

Parameter estimates from the logistic regression

Variable	Value	Estimate	Std. Error
Intercept		0,25	0,12
Level of Education	Short education	0,88	0,09
Level of Education	Middle education	1,27	0,09
Level of Education	Long education	1,85	0,10
Level of Education	Not recorded ^a	0,00	.
Sector of Employment	Private industry	0,47	0,06
Sector of Employment	Private non-industry	0,46	0,05
Sector of Employment	Government	0,71	0,05
Sector of Employment	Other	0,00	.
Living alone status	Living alone	-0,72	0,05
Living alone status	Not Living alone	0,00	.
Urbanisation	G011	-0,59	0,10
Urbanisation	G012	-0,47	0,10
Urbanisation	G013	-0,36	0,12
Urbanisation	G014	-0,23	0,14
Urbanisation	G021	-0,17	0,10
Urbanisation	G022	-0,44	0,11
Urbanisation	G023	-0,23	0,10
Urbanisation	G024	-0,31	0,11
Urbanisation	G031	-0,30	0,11
Urbanisation	G032	-0,03	0,10
Urbanisation	G033	-0,11	0,12
Urbanisation	G034	0,00	.
Employment status	Unemployment ^b	-0,24	0,04
Employment status	Employment ^b	0,00	.
Marital status	Not Married	-0,21	0,05
Marital status	Married	0,00	.

^aLevel of education was not recorded. Mostly elderly people with short education.

^bWas registered as employed/unemployed the quarter before .

10. Urbanisation codes are :

G011- G014 Greater Copenhagen incl. suburbs (whole metropolitan area).

G021- G024 Provincial municipalities with cities of min. 10.000 inhabitants.

G031- G034 Rest of the country.

Within each group, increasing numbers indicate decreasing urbanisation.

V. INTERPRETATION OF THE PARAMETER ESTIMATES FROM THE LOGISTIC REGRESSION

11. The interpretation of the parameters used is as follows:

Education: For all the four groups, the more education the higher response rate.

Sector: People working for the government had higher response rates than persons working in the private sector. People working for the private sector had higher response rate than non-working persons.

Living alone: People living alone have a relative low response rate.

Employment: Unemployed people also have a relative low response rate.

Urbanisation: The metropolitan area has a lower response rate than provincial municipalities.

Marital status: Married people have a higher response rate than non-married people.

VI. INCOME STRUCTURE

12. We have used a normal linear regression model to find the variables, which can explain the income. Only income data from the interviewed respondents have been used in the estimations. Income does not follow a normal distribution, but the regression is only used to identify the factors, which explains most of the variation . The following factors explained income (in order of decreasing importance):

Type 3-tests from the linear regression

Source	DF	F Stat.	Pr>Chi-Sq.
Sector of employment	3	779	0,0001
Age group	5	574	0,0001
Level of Education	3	471	0,0001
Sex	1	748	0,0001
Employment status	1	298	0,0001
Marital status	1	84	0,0001

Parameter estimates from the linear regression

Variable	Value	Estimate 1000DKK	Std. Error 1000 DKK
Intercept		125,9	4
Sector of Employment	Private industry	90,4	2,2
Sector of Employment	Private non-industry	85,8	2,0
Sector of Employment	Government	49,5	2,0
Sector of Employment	Other	0	.
Age group	15-19 years old	-142,3	3,8
Age group	20-29 years old	-51,7	2,9
Age group	30-39 years old	0,0	2,7
Age group	40-49 years old	14,2	2,8
Age group	50-59 years old	10,4	2,8
Age group	60-67 years old	0	.
Level of Education	Short education	16,5	3,4
Level of Education	Middle education	38,2	2,5
Level of Education	Long education	88,7	3,7
Level of Education	Not recorded ^a	0	.
Sex	Woman	-36,9	13
Sex	Man	0	.
Employment status(register)	Unemployed ^b	-24,7	1,4
Employment status(register)	Employed ^b	0	.
Marital status	Not Married	-13,7	1,5
Marital status	Married	0	.

^aLevel of education was not recorded. Mostly elderly people with short education.

^bWas registered as employed/unemployed in previous quarter.

VII. INTERPRETATION OF THE PARAMETER ESTIMATES FROM THE LINEAR REGRESSION

13. The interpretation of the parameters used is as follows:

Sector of Employment: People employed in the private sectors earn 90,000 DKK more per year than the unemployed. People employed by the government earn 50,000 DKK more than the unemployed.

Age group: The 15-19 years old and the 20-29 years old have a relative low income as expected.

Level of Education: The higher education gives higher income.

Sex: Women earn 37,000 DKK less than men, which can only partly be explained by the relative higher number of part-time working women.

Employment status: People unemployed in the last quarter earn 24,700 DKK less.

Marital status: Married people earn 13,700 DKK more than unmarried people.

VIII. IMPUTATION

14. The following imputations were tested:

- i) Group mean imputations with no grouping variable.
- ii) Group mean imputations with Sector of Employment as grouping variable.
We would have tried group mean imputations with more variables if SOLAS allowed it.
- iii) Hot Deck imputations with Sector of employment, Age group, Level of Education and Sex as variables for sort.
- iv) Multiple Imputations is tested with Sector of Employment as grouping variable and Level of Education, Living alone status, Employment status (register), Urbanisation as possible covariates.

15. Sector of Employment was used as grouping variable because that variable was found to have the largest correlation with income. Level of Education, Living alone status, Employment status (register) and Urbanisation were used as possible covariates because they explained the missing data structure best. We also would have liked to have used Employment status and Marriage status as possible covariates, but the computer crashed.

IX. RESULTS

16. In the following table, means of imputed income values in each sector of employment are compared to the actual value of income.

Sector of Employment:	Private industry	Private non-industry	Government	Not employed	All Sectors
Type of Imputation					
Actuarial before taxes income	201554	186191	113313	104006	143082
Group mean imp. no group variable	173632	173632	173632	173632	173632
Group Mean imp. group by Sector	218380	212844	146970	118709	166118
Hot deck Imputations	184999	186474	119708	119110	146850
Multiple Imp. no. 1	213455	204739	121918	112597	154654
Multiple Imp. no. 2	209460	204543	117245	115504	153599
Multiple Imp. no. 3	209853	204532	122122	115250	154889
Multiple Imp. no. 4	211871	210659	120104	116504	156577
Multiple Imp. no. 5	213990	209604	115836	112472	154254
Multiple Imp. average	211726	206815	119445	114465	154795

17. Group Mean Imputations with no grouping variable has the largest bias and is therefore worst. There is surprisingly little difference between the results from the three other methods. It is hard to see from this table which imputation method is best.

18. Root mean square of the difference between actual income and imputed income values in for each sector of employment.

Type of Imputations	Private industry	Private non-industry	Government	Not employed	All Sectors
Type of Imputation					
Group mean imp. no group variable	104862	106581	109956	86058	101153
Group Mean imp. group by Sector	102467	109143	97902	52672	90179
Hot deck Imputations	113168	143711	80280	64692	101254
Multiple Imputation no. 1	147672	153480	123032	75738	123764
Multiple Imputation no. 2	130852	150884	123088	76613	119879
Multiple Imputation no. 3	141783	148246	115396	77831	119409
Multiple Imputation no. 4	144194	163184	116156	79379	125010
Multiple Imputation no. 5	137712	160810	117438	75909	122626
Average of Multiple Imputation 1 -5	140442	155321	119022	77094	122138
Average of the values imputed for Multiple Imputation 1-5	110130	115177	87065	57798	91556

19. From the table above it can be seen that the imputed income produced by group mean imputation has the smallest deviation from the true values. By looking at "Average of the values imputed for Multiple Imputation 1-5" it can be seen that SOLAS performs well.

References

Roderick J. A. Little & Donald B. Rubin [1987]: "Statistical Analysis with Missing Data ": John Wiley & Sons Inc.

Donald B. Rubin [1987]: "Multiple imputation for non-response in surveys " John Wiley & Sons Inc.

R. Kozak, from Generalized System Method Section BSMD December 22, 1998 "Solas for Missing Data Analysis. Software Evaluation."

SOLAS for Missing Data Analysis 1.0 - User Reference. Statistical Solutions 1997.