DFG Project FOR 756 “Vulnerability in Southeast Asia”

Household survey - Second wave (2008)

Data Cleaning Guidelines

Version 1.3

09 September 2008

Table of contents

1. Introductory remarks

1. During the data cleaning process the following different versions of cleaned data are generated:
    i. Version 1.0 – raw data
    ii. Version 2.0 – cleaned data excluding income and consumption relevant variables
    iii. Version 2.1 – cleaned data including income and consumption relevant variables

    Further versions will be generated if necessary.

2. Bernd Hardeweg coordinates the data cleaning process. This includes
    i. the creation of a data cleaning specific mailing list;
    ii. the composition of weekly updates on the data cleaning progress that are sent to all relevant stakeholders; and
    iii. the follow up on deadlines.

3. The software of choice for data cleaning is STATA. It is worked with long-format data.

4. The guiding principle of the data cleaning process is to change original data as little as possible. In other words, observations that are outliers at first sight are not automatically to be dropped or replaced. Plausibility checks (see paragraph 14) may reveal that outlying values are actually plausible and, consequently, can be kept. This procedure is in line with Deaton and Zaidi (1999, p.25) who state that it might be "unclear whether the 'outlier' is genuine or not" and that "the analyst must make a judgement that balances the desirability of keeping any reasonable number [of observations] against the risk of contaminating the aggregate."

2. Version 1.0 – raw data

5. If not stated otherwise, work related to the generation of version 1.0 is done by Hannover (agricultural economics).

6. Version 1.0 contains “raw” data. Excluded from version 1.0 are
    i. identical rows that have been traced down to erroneous double entry (e.g. the same insurance);
    ii. rows with information from the first wave – such as land size – which have been filled in by enumerators in the field and do not correspond with newly collected data; and
    iii. household members listed in the first wave who the respondent does not consider as members even for the first wave, i.e. erroneous prior information.

7. Included in version 1.0 are newly created, household specific row IDs in all cases where no explicit row identification is possible via information form the questionnaire. That is, the risk section (3.2; page 23 of the questionnaire), for example, will not contain any row IDs since all observations can clearly be identified via the event ID. By contrast, the agricultural section (4.2; page 29 of the questionnaire) will contain newly created row IDs.

8. All monetary values are given in purchasing power parity adjusted US dollars. The common unit for land size is hectare.

9. To “90 (others, specify)”-observations a self-contained category is assigned if the same / very similar specifications of “90” cover at least one percent of the total number of observations. The respective recoding of categories is implemented section specifically by the different sub groups.

3. Version 2.0 – cleaned data excluding income and consumption relevant variables

10. Data cleaning starts simultaneously in all sections. However, in order to conduct plausibility checks (cf. paragraph 14) it might be necessary to refer to information from other sections. Therefore, data is cleaned according to the following causality chain:
    i. Sections 1, 2, and 3
    ii. Sections 4, 5, 6, and 9.2
    iii. Sections 7, and 9.1
    iv. Section 8

    Since sections 1, 2, and 3 are at the beginning of the chain, there are tighter deadlines for these sections than, for example, for sections 7, and 9.1:

11. Deadlines (and responsibilities; cf. appendix A):[1]
    i. Sections 1 (agricultural economics; Hannover), 2, and 3 (development economics; Göttingen) are cleaned by Friday, 26th of September;
    ii. Sections 4 (agricultural economics; Hannover), 5, 6 (economic geography; Hannover and Gießen), and 9.2 (development economics; Göttingen) are cleaned by Friday, 10th of October;
    iii. Sections 7 (finance; Hannover and Frankfurt), and 9.1 (agricultural economics; Hannover) are cleaned by Friday, 17th of October;
    iv. Section 8 (development economics; Göttingen) is cleaned by Friday, 24th of October.

12. Treatment of missing values:
    i. "97 (don't know)", and "98 (no answer)": judgement: if the observation reflects a value, "97" / "98" are replaced via the standard replacement procedure (cf. paragraph 15). If the variable reflects a code, it is checked whether a replacement is necessary for any analysis. If a replacement is considered to be necessary, the observation is replaced by the most plausible category (e.g. by unit "kilograms" if quantity with unknown unit is similar to quantities in kilograms). In any case 97 and 98 as an indication of missing values should

[1] Version 2.1, i.e. the version including income and consumption relevant variables, will be cleaned by the end of November. Hannover (agricultural economics) is responsible for the income aggregate. Göttingen (development economics) is responsible for the consumption aggregate.

be replaced by either valid data or .a (for 97) or .b (for 98) in STATA[2]. This avoids erroneous inclusion of these numbers in calculations.

ii. "99 (does not apply)": For ratio and interval scaled variables, the respective sub-project in charge of a given section decides whether to replace 99 with zeroes. As a guideline, (monetary) values can often be replaced by zero (e.g. does not apply for education expenditures means that these have been 0). On the other hand, prices should usually not be replaced by 0, because averages calculated over valid observations and zeroes will be affected by such replacement. "99" in coded variables is to be replaced by ".", the STATA system missing value.

iii. Missing observations (e.g. no information on education expenditures in section 8) are treated analogously. Depending on the purpose of analysis, missing values can be replaced by "0". If the variable reflects a code, it is checked whether a replacement is necessary for any analysis.

13. Identification of outliers:

i. The standard procedure for the identification of outliers is to calculate lower and upper bounds by adding and subtracting, respectively, two standard deviations from the median of any group with at least ten observations (e.g. food expenditures; groups with less observations, e.g. consumption of fruits in units of quantity that cannot be transformed into kg, are checked by hand). Values below and above these bounds are considered to be outlying.

ii. Deaton and Zaidi (1999, p.25) suggest "to do this [search for outliers] in logs as well as in levels". We follow this approach and apply the standard procedure to logs and levels before taking a closer look at the outlying values. That is, if the analysis in logs yields ten outliers and the one in levels thirteen, whereof four have not been identified in logs, ten plus four observations are checked.

iii. Whenever the standard procedure yields negative lower bounds but negative values are not possible (e.g. in the case of home consumption), the lower bounds are replaced by zero.

iv. The plausibility of lower and upper bounds is constantly checked.[3]

v. The standard outlier analysis just serves as an indication of observations that might be replaced.

14. Plausibility checks:

[2] You can use the command mvdecode varlist, mv(99=. \ 98=.a \ 97=.b)

[3] For example, if the median equals zero the upper bound is likely to be implausibly low.

i. Outliers identified by the standard procedure are analyzed with regard to their plausibility. Only outliers that are considered to be implausible observations are replaced.

ii. Every sub group decides individually on the design and complexity of its plausibility checks. However, every plausibility check starts by examining whether the notes to the questionnaire (section 1) or remarks for plausibility violations provide relevant explanations.

iii. In the context of plausibility checks data consistency is also scrutinized. For example, in the livestock and aquaculture section (4.3.1; page 35 of the questionnaire) it is examined whether the stock at the end of the year equals the stock at the beginning of the year plus (minus) additions (disposals).

15. Replacement of missing and outlying values:

i. For the replacement a common STATA code that can easily be applied to different variables, as well as to possible future waves is used. The code will be provided by Tobias Lechtenfeld.

ii. The standard replacement procedure uses local means. It is used, for example, to impute market prices for certain goods. This approach follows exactly the replacement procedure applied in the context of the income (and consumption) aggregate of the first wave: "In most cases the mean of each variable with sufficient cases and plausible information for the nearest possible level of sampling (village, commune, district and province) was used [for replacement]. As a threshold a minimum of five cases was introduced." (DFG FOR 756, 2008, p.2). The proceeding might be amplified by certain “by” and “if” conditions (e.g. replace value with mean wage of same (i.e. “by”) occupation and gender if occupation counts as off farm employment).

iii. The “advanced” procedure uses local means, as well as location and household characteristics in order to estimate replacement values. In this context, education expenditures might be, for example, regressed on mean expenditures for education in the same commune (i.e. local means), the number of currently enrolled household members (differentiated by primary, secondary, and tertiary education; i.e. household characteristics), and the distance to the nearest school (information from last first wave’s village head questionnaire; i.e. location characteristics). In this case, implausible outliers would be replaced by values as predicted by the regression results.

iv. Sub groups decide individually and on a case by case basis which replacement procedure is more reasonable / to be used.

16. Generation of new variables:

i. New variables that contain cleaned, i.e. “workable”, observations are generated. These variables are marked with an “x” at the second digit (e.g. _x43202 instead of __43202). The original observations are only kept in version 1.0.

ii. No dummies identifying treated observations are generated. Whether an observation has been cleaned can easily be deducted from the difference between the values of the new and the original variable.

iii. As in the first wave, variables that facilitate research are generated (e.g. “__21022 (nucleus household membership – yes=1/no=0”). Variables that have to be generated in this context are listed in DFG FOR 756 (2007) where they are marked with a “D” or “I” in column “Src*”.

iv. Consumption and income relevant variables are generated when the consumption and income aggregates are imputed, i.e. within the scope of version 2.1.

17. Documentation of data cleaning:

i. Every sub group composes a document that keeps record of all the steps which were taken during the data cleaning process. Subsequently, all documents from the sub groups are merged to a handbook that is published on the project’s web site.

ii. In the STATA do-files every command is explained. All do-files are attached as appendices to the documents and the handbook, respectively.

iii. Every document presents section specific key facts that reflect for each variable

   a. the total number of observations;
   b. the number of outlying observations (which have actually been treated);
   c. the number of missing observations;
   d. the number of replaced observations;
   e. the number of dropped observations; and
   f. the number of households with treated observations.[4]

Additionally, outlying, missing, replaced, and dropped observations are stated as percentage share of the total number of observations. In the case of households, the number of households with treated observations is given as

[4] Treated observations are replaced or dropped ones.

percentage share of the total number of households included in the section.[5] The key facts are presented as shown exemplarily in appendix B.

Bibliography

Deaton, Angus and Salman Zaidi (1999), *Guidelines for Constructing Consumption Aggregates For Welfare Analysis*, Working Papers 217, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton.

DFG FOR 756 (2008), Description of the procedure for replacement of missing values and other changes in the dataset - Household survey 1st wave 2007, version from 16.03.08, University of Hannover.

DFG FOR 756 (2007), Data dictionary - HH Survey FOR 756, University of Hannover.

Appendices

Appendix A – Deadlines and responsibilities		
Section	**Deadline for data cleaning**	**Responsible for data cleaning**
1	Friday, 26th of September	agricultural economics; Hannover
2	Friday, 26th of September	development economics; Göttingen
3	Friday, 26th of September	development economics; Göttingen
4	Friday, 10th of October	agricultural economics; Hannover
5	Friday, 10th of October	economic geography; Hannover and Gießen
6	Friday, 10th of October	economic geography; Hannover and Gießen
7	Friday, 17th of October	finance; Hannover and Frankfurt
8	Friday, 24th of October	development economics; Göttingen
9.1	Friday, 17th of October	agricultural economics; Hannover
9.2	Friday, 10th of October	development economics; Göttingen

Appendix B - Key facts from … section						
Variable	Total number of observations	Number of outlying observations (% of total)	Number of missing observations (% of total)	Number of replaced observations (% of total)	Number of dropped observations (% of total)	Number of households with treated observations (% of total)
…	…	…	…	…	…	…

[5] At the beginning of every section specific documentation the total number of households included in the respective section is stated.