Guidelines for Data Analysis

Arne Henningsen

Department of Food and Resource Economics, University of Copenhagen
arne.henningsen@gmail.com http://www.arne-henningsen.name/

23rd August 2016

I wrote these guidelines in order to have a handy summary of the most important hints for conducting data analysis both for myself and for others, e.g. my students. These guidelines intend to make data analysis more reliable and to ensure reproducibility, which is a key principle of scientific research. The hints in these guidelines have been collected from various sources and/or are based on my own experience as applied econometrician and supervisor of MSc and PhD theses. For instance, several points of these guidelines were taken from or were inspired by Sandve et al (2013) and Leek (2015). As I want to keep these guidelines as short as possible, I only included those points that I (subjectively) find most important. More details are available in, e.g., Sandve et al (2013) and Leek (2015) and are taught in the "Data Science Specialization" of Johns Hopkins University that is available at Coursera (https://www.coursera.org/specializations/jhu-data-science). Following these guidelines may require some learning, which could slow down the data analysis in the beginning. However, the learning does not take a long time and it will significantly increase your productivity so that you will soon have made up the time that you used for learning in the beginning. On top of this, you will have more reliable and reproducible results. Feedback on these guidelines is highly appreciated.

General data handling and reproducibility

1. You have saved your raw data and you do **not** modify them (no matter whether your raw data are in electronic form or in non-electronic form).

2. If your raw data are in non-electronic form (e.g. written on paper questionnaires), you enter them to an electronic database (e.g. a spreadsheet software) without making any manual modifications so that the electronic version of your raw data is an identical copy of the non-electronic version of your raw data.[1]

3. You do **not** manually modify any of your data files.[2]

4. If necessary, you tidy your data set so that each variable is one column and each observation is one row. If you do this, you do **not** do this manually but use script files for doing this.

5. You save the tidied version(s) of your data set.

6. You clean your data (e.g. correcting data errors, removing outliers) by using script files and you do **not** do any manual cleaning.

7. You save the cleaned / processed versions of your data set.

8. You use script files to do all calculations and to perform the (final) analysis. (Only the exploratory data analysis can be done manually.)

9. You have your data files and script files in a version control system (e.g. Subversion, Git) in order to track (intended and unintended) changes.

10. If your analysis involves some element of randomness, your script files set the "random seed" of the pseudo random number generator so that the script gives **exactly** the same result every time when it is executed. Finally, you modify the random seed(s) and re-run your script(s) a few times in order to assess the robustness of your results with respect to different random seeds.

11. You note the exact names and versions of all software packages that you use in your analyses (or even better: you archive them), because different (e.g. newer) versions may give different results or do not work without modifying the script files.

12. You ask someone else to run your analysis (starting from the raw data) and they get the same results.

13. You use script files to generate all tables and figures for your publications so that these tables and figures are automatically updated if you change anything in your analysis.

[1] If you think that there is a mistake (e.g. a typo) in your non-electronic raw data, you do **not** try to correct this (supposed) mistake while entering the data, because this manual modification would not be reproducible and your electronic raw data would no longer be an identical copy of your non-electronic raw data. In this case, you should make a note so that you will later remember to address this (supposed) mistake (see point 6). If you later notice that the electronic version of your raw data is not an identical copy of your non-electronic version of your raw data (e.g. because somebody made a mistake when entering the data), you can use a script file to correct this mistake or—if you have the electronic data file in a version control system—correct this mistake manually and clearly document this correction in the log message of the version control system.

[2] Perhaps with one very limited exception as explained in footnote 1.

14. You make your script files, the resulting output files, your notes about software packages and their versions, and if possible also your data available to co-authors, supervisors, collaborators, reviewers, and the general public, because this signals a high quality, trustworthiness, and transparency of your analysis.

15. You have created a flowchart that illustrates the relationships between all data files, script files, and output files (e.g. log files, tables, figures). An example is given in Figure 1. There is no circular relationship between any files.

Checking the data and exploratory analysis

1. You make sure that missing values are explicitly coded as missing values (**not** as "0", "–1", "999", or something else) and zero values are explicitly coded as zero values (**not** as missing values).

2. You check all variables for missing values.

3. You check the values and units of all data points and make sure that they are in the right range.

4. You use appropriate univariate plots (e.g. histograms or density plots for continuous variables, bar plots for categorical variables) to check every variable in your data set.

5. You check all variables for outliers.

6. You use appropriate bivariate plots (e.g. scatter plots, boxplot diagrams, spine plots) to check the relationship between various pairs of variables.

7. When variables are considerably (right-) skewed, you plot them in log scale.

8. You check that pairs (or subsets) of variables do not contain contradictory or implausible information (e.g. professional experience larger than age, total labour income divided by total hours worked gives an implausible wage per hour).

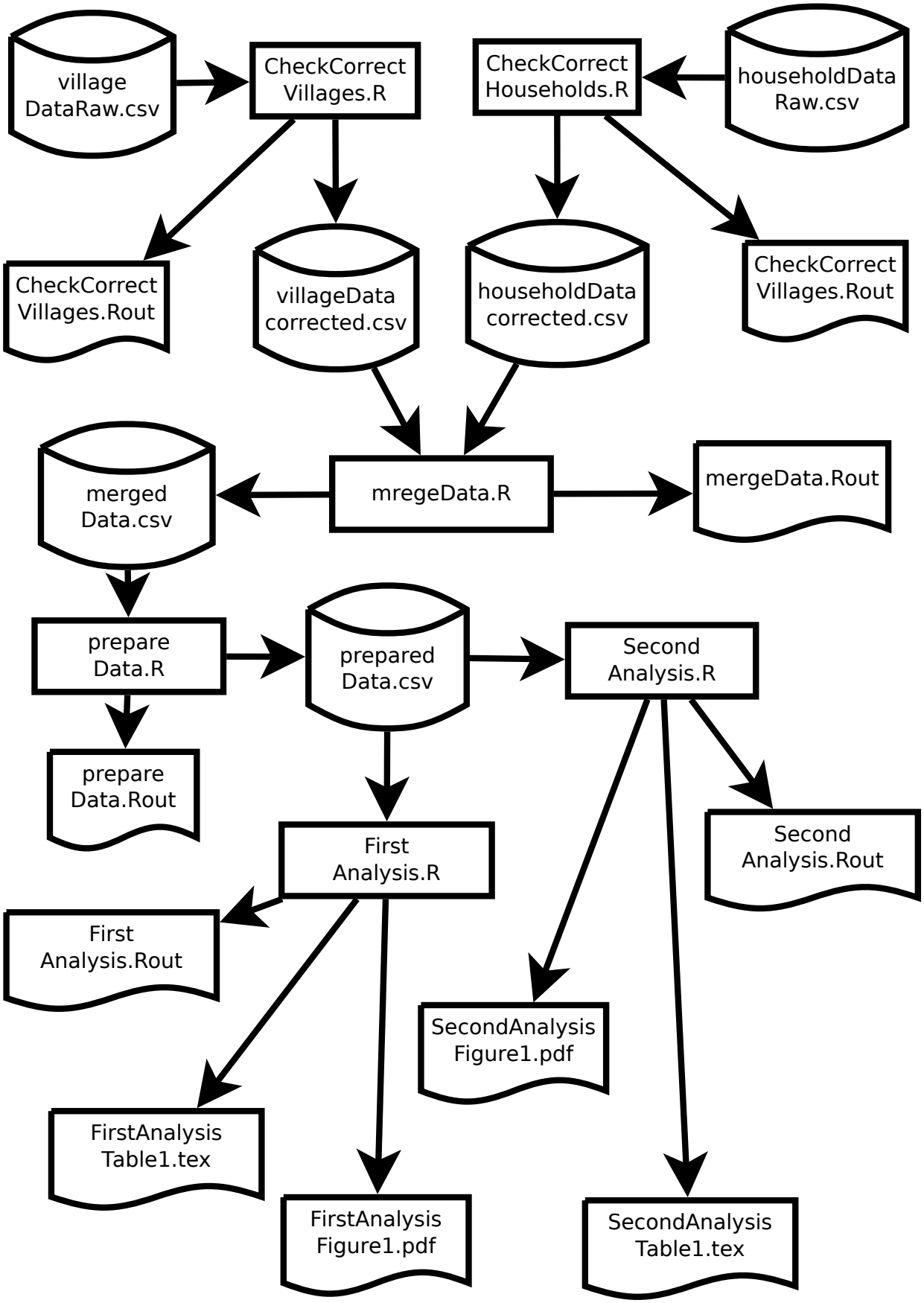

Figure 1: Flowchart that gives an example of a relationship between data files (cylinders), script files (rectangles), and output files (rectangles with a wavy base)

Appendix A: Preparing a questionnaire for a survey

I have seen many (mostly young) researchers who started their research by developing a questionnaire and conducting a survey. After the survey, they tried to figure out, which models they can estimate with their data. Finally, they searched for research questions that can be answered based on the estimates of their models. However, these steps should be done in the opposite order:

1. define very specific research questions;
2. define detailed specifications of the models that you want to estimate in order to answer the research questions defined in the first step;
3. list all (dependent, explanatory, control, instrumental, ...) variables that you need to estimate the models defined in the second step;
4. describe how you operationalize the variables listed in the third step;
5. formulate the questions for your questionnaire;
6. conduct a pre-test of your survey;
7. enter the data from the pre-test;
8. check the data obtained in the pre-test and use these data to conduct **all** analyses that you plan to conduct with the data from the main survey; do the testing and the analyses as described in the main part of these guidelines;
9. revise the questionnaire if necessary, e.g. modify questions that frequently result in missing, implausible or unreliable values and add questions to obtain information that is required to conduct the planned analyses but that is not available from the answers to the pre-test questionnaire; and
10. repeat the pre-test in case of substantial changes of the questionnaire.

Furthermore, the questionnaire must be designed in a way that the obtained data set distinguishes between zeros and missing values.

Appendix B: Working with RStudio and Subversion (SVN)

You never safe the R workspace; if you have accidentally saved it, you remove the saved workspace (i.e. the file ".RData").

After each individual 'task' (even if it took you just 5 minutes), you 'commit' your changes to the Subversion repository. In general, the more frequently you commit your changes, the higher the quality and reliability of your work will get and the more productive you will get (at least in the medium and long run).

Each time before you 'commit' any changes to the Subversion repository, you do the following things:

1. You check whether the R script can be 'run' / 'sourced' without manual intervention in a 'fresh' environment, i.e.
   a) you 'clear' all objects from the R workspace, e.g. by clicking on the broom symbol in RStudio's "Environment" tab or by choosing in RStudio "Session" → "Clear Workspace...,"
   b) you restart R, e.g. by choosing in RStudio "Session" → "Restart R,"
   c) you click on "Source" in the upper right corner of your R-script in RStudio, and
   d) you check if there are any *error* messages or any *warning* messages that indicate a problem. If there are *error* messages or *warning* messages that indicate a problem, you fix them and start over at step 1a. There may be a few specific unavoidable *warning* messages that could be ignored if they do not indicate a serious problem, but *error* messages can never be ignored.
2. You use the "diff" feature of the Subversion system ("Show Changes" in SmartSVN) to check whether all modifications / changes are indeed intended and make sense. Otherwise, you can manually revert individual changes in RStudio or you can use the "revert" feature of the Subversion system ("Revert" in SmartSVN) to revert *all* modifications that you did since your previous 'commit.'

This will greatly increase the quality and reliability of your work and in the medium and long run save you a lot time, i.e. make you more productive :-)

References

Leek J (2015) The Elements of Data Analytic Style. Leanpub, URL http://leanpub.com/datastyle

Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten simple rules for reproducible computational research. PLoS Computational Biology 9(10):1–4, DOI 10.1371/journal.pcbi.1003285