Fall 2023

DIME Analytics

Reproducible Research Fundamentals

September 25-29, 2023

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Fall 2023

DIME Analytics

Cleaning Data Hands On

Benjamin Daniels

DIME Analytics

bdaniels@worldbank.org

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Fall 2023

DIME Analytics

Cleaning Data Hands On

During the training, find all materials in our shared OneDrive here

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Comparing Primary and Secondary Data

Primary Data

- Custom-made
- Current insights*
- Time-consuming
- Can be expensive

Secondary Data

- Economical
- Broad database
- Potential misalignment
- Quality concerns
- Immediate

Data Exercise

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Exercise

Apply the tasks you've learned in the last few sessions to this data set:

Data cleaning: Clean the dataset using a script or do-file.

1. Check for data collection metadata variables not needed for analysis and drop them (id test data)
2. Make sure there is one or more identifying variables in the data
3. Make sure each variable has a correct data type
4. Handle missing values appropriately using commands like 'codebook, compact'
5. See if there are any special characters in the data and remove them.
6. Check that all variables have a label in the working language of your team (assume it's English for this exercise)

Exercise (continued)

Documenting metadata:

Create or export a codebook or data dictionary of your cleaned dataset

Documenting data cleaning and consistency:

Document all the data cleaning tasks and the changes you apply to the dataset

Review all the variables and check that they have consistent values. Document your checks and any anomalies you consider important for next stages.

Importance in Social Research

Benefits of Integration:

- Richer analysis
- Enhanced reliability
- Cost-effective

Tailoring the Cleaning Process:

- Ensures data reliability
- Prevents misinformation
- Facilitates valid conclusions

Data Cleaning Tips

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Data Cleaning Tips

Commands for testing that a variable is uniquely and fully identifying:

isid

levelsof

codebook

labelbook

Commands for codebook creation and management, and labelling:

iefieldkit

iecodebook template

iecodebook apply

iecodebook export

Don’t forget to use [help date] for managing dates and times

Saving clean data

- During the data cleaning process, you might have saved multiple intermediate files, for example if you cleaned long modules separately to make your code more readable
- After cleaning your data and merging it back together, you'll want to save a final cleaned data set, containing all variables you will use in the analysis
- This new data set will probably be quite heavy. Use [compress] to save your variables in the most economic format

Naming files

- Make sure all output files, datasets and others are clearly and uniquely labeled, i.e.: “desc_stats_tmt_only.xls” “input_plan_adm_data.dta”
- It’s often desirable to have the names of your data sets and do-files linked, so it is easy to understand which do-files is creating which data set, such as “merge.do” and “merged.dta” or “cleaning.do” and “clean.dta”
- Do not use v1, v2 etc. for any final files. This leads to bugs in do-files that depend on these files when a new versions is added.

Hints for metadata documentation

Variable labels must be short and self-explanatory, as they will be used in tables and graphs

However, there is much more information that is useful for someone opening the data for the first time

This information should be stored in a data dictionary/codebook, including:

- The definition of each variable or corresponding survey question
- The number of missing observations in each variable
- Summary statistics
- Any field notes or corrections made to each variable

Cleaning documentation and data consistency

Documenting data cleaning

Describe in order the data cleaning tasks you're doing. Use the working language of your team

Even if you don't edit the dataset after a task (for example, there might not be duplicated entries in your data), it's a good practice to document the task and note that no changes were implemented

Check variables consistency

Check that values are consistent across variables

For example, if an individual is male, then he cannot be pregnant

This kind of inconsistency should usually be corrected during the high-frequency checks, but often times there's no time when the enumerators are in the field to identify and correct all of them

So if you find any issues, create flag variables that identify observations with inconsistent values

Fall 2023

DIME Analytics

Cleaning Data Hands On

Benjamin Daniels

DIME Analytics

bdaniels@worldbank.org

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT