Things to Do and Think About When Cleaning Data

MMN 10-10-2012

Goal is to make a well-labeled, well documented file with internal inconsistencies resolved that stands on its own and that you would be happy giving someone else to use.

Assume that the cleaning do file will be given to the agency, so keep it tidy and well-organized. Indent consistently, mark queries and unresolved issues, and comment well but succinctly.

Add a discussion of data/project history, precursor and successor data and program files, and outstanding issues to the beginning of the file. Try to keep the discussion updated.

Think in terms of writing code that someone at SDP or in the agency can use as a model when working on other agencies or updating datasets with additional years of data.

Don’t drop records unnecessarily or restrict samples at the cleaning stage. Try not to lose information.

Add dataset or variable notes when restructuring data, dropping records, or changing variable values.

Derived variables that are needed for analysis but where definitions are open to discussion should not be added at the cleaning stage. They should be defined in one place after files are merged so that they can be referenced easily.

Ideally, key structure and counts should already be resolved, so that you only have to load the file and quickly check the uniqueness level.

Run codebook. Delete empty variables, and systematically check all variables with missing values, documenting your findings.

Tab categorical variables, and do summ, detail for numerical variables, checking for reasonableness.

Cross-tab variables by school year, etc. to check data coverage and consistency.

Cross-tab related variables by each other (including missings), to the extent time permits.

When codes and descriptions are in separate variables, check to see if the relationship between them is one to one. Use descriptions as labels for codes when possible.

Convert date variables to Stata date format, and check for out of range values.

Encode categorical variables as necessary, and define and label values.

Label school codes and district codes with school names and district names.

Label variables and values using upper case title conventions.

Label school years as fall school year or spring school year as appropriate.

Resolve conflicts/inconsistencies at the record level for variables like gender and ethnicity, choosing modal, min, max, or random values as appropriate.

If you have documentation or codebook information from the agency or the internet, add the information to the file. Ideally, this research has already been done. Typical categories of information worth adding include enrollment, withdrawal and graduation codes, school names and types, job code structure and descriptions, class subjects and descriptions, test performance codes, salary lanes and steps, degree and certification codes, school year beginning and end dates…

Define time series structure (tsset) for panel data, and do progression checks if time permits. If you don’t know how the agency defines experience, mark it as a query to be raised later.

Typical items that have to be researched in more depth and separately from data cleaning and processing are information about schools and information about tests.

The perfect is the enemy of the good; do the best job possible with time available.