How to Search Datasets for Personally Identifiable Information

May 10, 2013

To ensure a Stata dataset does not contain personally identifiable information (PII), you should carefully review the variables it contains: it may not be immediately clear that a variable is PII. However, sometimes it is useful to complete a first sweep of one or more datasets for clear instances of PII.

One way to do this is the `lookfor` command in Stata. It searches all variable names and labels in a dataset for one or more keywords. For example, `lookfor name` lists all variables whose name or variable label contains the string `name`. In this case, a variable named `fname` (for "first name") would be listed because `name` is a substring of `fname`. `lookfor` also stores the list of variables in the saved result r(varlist).

To quickly search more than one dataset, use the `lookfor_all` command, available on SSC. To install, type `ssc install lookfor_all` in Stata.

Below is a list of keywords to consider searching for. The list is not exhaustive, and you may find other PII examples in the data security manual.

- `name`
- `birth` (to find variables related to the respondent's birthdate)
- `phone`
- `district`
- `county`
- `subcounty`
- `parish`
- `lc` (to find variables related to the respondent's "local council," a geographical unit in some countries)
- `village`
- `community`
- `address`
- `gps`
- `lat` (to find variables related to latitude)
- `lon` (to find variables related to longitude)
- `coord` (to find variables related to GPS coordinates)
- `location`
- `house`

How to Search Datasets for PII

To ensure a Stata dataset does not contain personally identifiable information (PII), you should carefully review the variables it contains: it may not be immediately clear that a variable is PII. However, sometimes it is useful to complete a first sweep of one or more datasets for clear instances of PII.

June 05, 2015