How do I document and search a Stata dataset?

Authored by

stats writer

July 1, 2024

To document and search a Stata dataset, you can follow these steps:

1. Start by opening your Stata software and loading the dataset you want to document and search.

2. Use the "describe" command to get a detailed overview of the dataset, including the variables, their labels, and value labels.

3. Next, use the "codebook" command to generate a codebook that provides a summary of all the variables in the dataset, along with their descriptions and other relevant information.

4. You can also use the "label list" command to display all the variable labels in the dataset.

5. To search for specific variables or values within the dataset, you can use the "find" or "search" commands. These commands allow you to search for specific strings, words, or values within the dataset.

6. You can also use the "lookfor" command to search for variables with specific keywords in their labels or descriptions.

7. Additionally, you can use the "tabulate" command to generate frequency tables and cross-tabulations to get a better understanding of the data in the dataset.

Overall, documenting and searching a Stata dataset involves using various commands and tools within the software to access and analyze the information contained in the dataset. This process can help you better understand and utilize your data for analysis and other purposes.

How do I document and search a Stata dataset? | Stata FAQ

When going through data management steps in Stata, there are several documentation features available that allow you to attach information to the dataset. This approach to documenting allows you to always have access to

your notes when looking at the data. Then, when working with the dataset, using some search commands can be an efficient way to find the desired variables for analysis and
avoid confusion.

First, we demonstrate how to add variable label, value labels, dataset notes,
and variable notes to data.

Adding a variable label and/or value label

A variable label allows you to describe the information contained in a
variable (thus allowing you to keep variable names more concise!). To add
a label to a variable, use the label variable command, then provide the
variable you wish to label and the label to apply. Below, we will create a
categorical variable called honsci based on the science variable.
Those with science scores of 60 or over are eligible for entry to a science honors
society. We will indicate this with a variable label.

```
use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear
gen honsci = (science >= 60)
label variable honsci `"honors eligibility"'
```

For purposes of data analysis in Stata, you will often need to code categorical variables as numeric. However, you can still have a string describing each numeric value in such a variable. Since the variable honsci contains 0 and 1, we will add value labels describing what each value represents. To do this, we will first define a label with the value descriptions using label define. Then we will apply the label to the values in honsci using label values.

```
label define elig 1 "eligible" 0 "not eligible"
label values honsci elig
table honsci
```

---

honors |
eligibility | Freq.
----------+-----------
not eligible | 153
eligible | 47

Adding variable or dataset notes

Notes in Stata are an excellent way to annotate your data. Notes can be
attached either to the dataset as a whole or to specific variables. Below,
we write a note to the dataset and a note to the ses variable.

note: Perhaps recenter score variables for analysis
note ses: This variable must be dummy coded for analysis

The command notes lists all of the notes in a dataset:

notes

_dta:

1. Perhaps recenter score variables for analysis

ses:

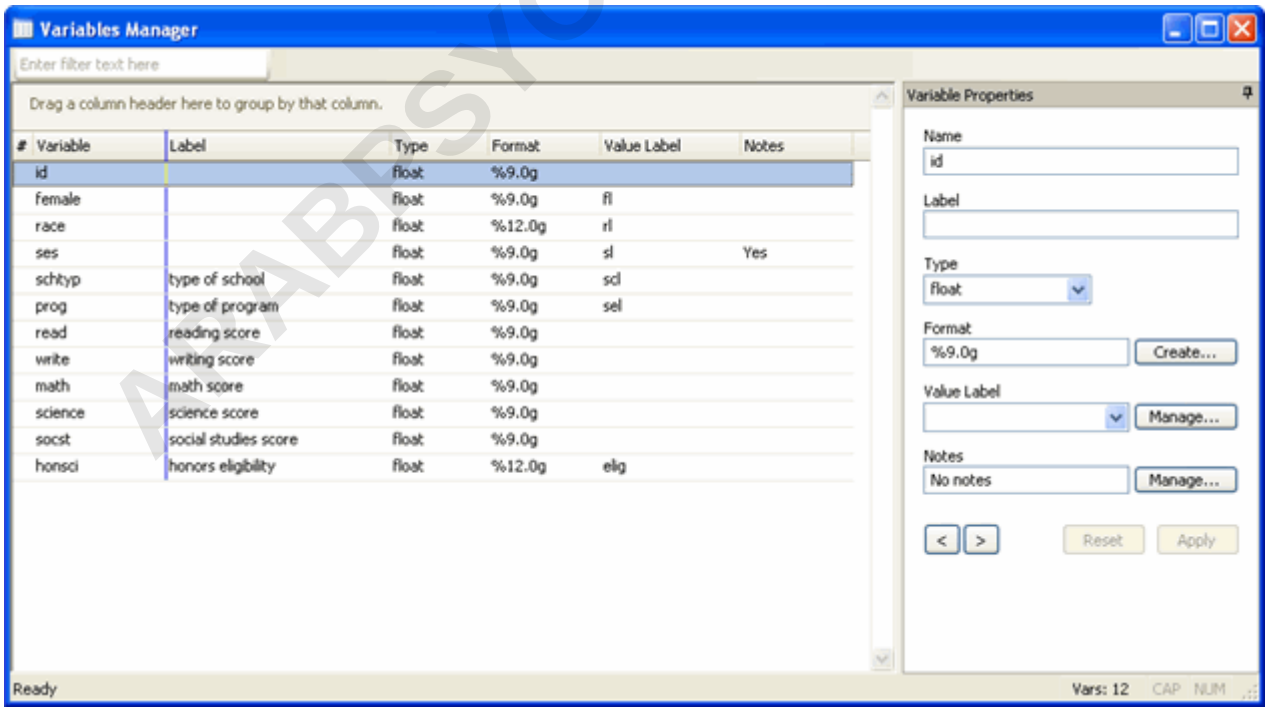
1. This variable must be dummy coded for analysis

All of these are included when a Stata dataset is loaded and all are searchable.

In Stata 11 and later, you can use the Variables Manager to see variable labels,

formats, and whether or not a variable has a value label or notes attached to

it.

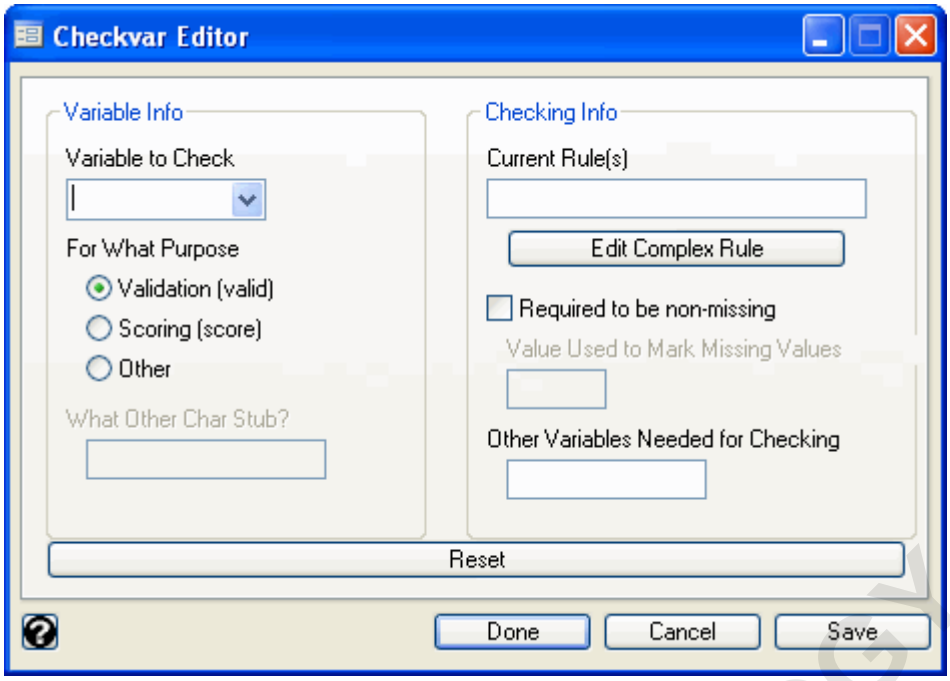


Adding data-validation checks

There is a package of Stata commands you can download, ckvar, that allows you to create "self-validating" datasets. For a full presentation of the package by its author, see Bill Rising's WCSUG07 slides.

We will briefly demonstrate how to add a rule to a variable and then check that the dataset does not have any observations that violate the accepted range. After downloading ckvar, you can use the ckvaredit to open a dialog window that allows you to enter validation rules for your variables.

ckvaredit



Our variable female should only take on values 0 and 1. We can select female from the "Variable to Check" list and then provide a rule for the acceptable range of values. We do this by entering in {0,1} in the "Current Rule(s)" box. When we click "Done", we see the equivalent code appear:

ckvareditSave female, stub(valid) req(0) rulechgflag(1) rule("in {0,1}")

We can check to see if there are any violations of rules by entering ckvar.

```
ckvar
Checking hsb2.dta on 28 Jan 2010 at 15:55:26:

There were no errors or missing required values!
```

If we wish to see what an error would trigger, we can change a value in our data and then rerun this check.

```
preserve
replace female = 3 if id == 1
ckvar

Checking hsb2.dta on 28 Jan 2010 at 15:58:36:

Variable name | Errors | Missing | Error-marker name
--------------+--------+---------+------------------
id | N/A | N/A | none
female | 1 | N/A | error_female
ses | N/A | N/A | none
schtyp | N/A | N/A | none
prog | N/A | N/A | none
read | N/A | N/A | none
```

write | N/A | N/A | none
math | N/A | N/A | none
science | N/A | N/A | none
socst | N/A | N/A | none
honors | N/A | N/A | none
awards | N/A | N/A | none
cid | N/A | N/A | none

restore

For more details on the types of rules that can be implemented, the linked presentation is the best source.

Searching in variable names and variable labels

If you are presented with a dataset and wish to find which variables have a
name or label containing a given string, you can use the lookfor command.
We will demonstrate this and similar commands using the small sample dataset,
hsb2, but it is likely more useful when you are looking at a dataset with
many variables or variables that are ambiguously named, but well labeled. First, we can

look for the string "male" among our variable names and labels.

lookfor male

storage display value
variable name type format label variable label

----
female float %9.0g fl

Stata returns a list of the variables that match our search-the variable
name "female" contains "male". Next, we can search for the string "score".

lookfor score

storage display value
variable name type format label variable label

----
read float %9.0g reading score
write float %9.0g writing score

math float %9.0g math score
science float %9.0g science score
socst float %9.0g social studies score

We see all of the variables with the word "score" in their label.

Searching in notes

You can search through these notes with the notes search command. We
can search for the string "analysis".

notes search analysis

_dta:
1. Perhaps recenter score variables for analysis

ses:
1. This variable must be dummy coded for analysis

Stata returns the notes containing the string and indicates if they are dataset notes or variable-specific.

Searching a directory (several datasets)

Often in an analysis, multiple datasets are analyzed. The lookfor_all command allows searches through all Stata files in the current directory. This is a user-written command and can be easily downloaded (see Stata FAQ: How do I use the search command to search for programs and additional help?).

Use the pwd command to learn Stata's current directory. You can change the working directory using cd or by choosing "Change Working Directory" from the "File" menu. The code below searches for the string "school" in the variable names and labels of the .dta files in the working directory.

```
pwd

D:Data

lookfor_all school
```

use "D:/Data/hsb2.dta"
variables: schtyp
use "D:/Data/hsbmis.dta"
variables: schtyp

Total 5 out of 5 files checked in "D:/Data/"

Stata returns the commands needed to load the datasets containing the
variable(s) of interest and indicates the names of the variables in each dataset
containing the string in its name or label.

Searching in value labels

To see all of the value labels present in a dataset and the list of variables
to which they have been applied, you can use the labelbook
command. Below, we demonstrate this (but show only one of the output labels to
save space).

labelbook

---

value label elig

---

values labels
range: string length:
N: 2 unique at full length: yes
gaps: no unique at length 12: yes
missing .*: 0 null string: no
leading/trailing blanks: no
numeric -> numeric: no
definition
0 not eligible
1 eligible

variables: honsci

Variables in Stata can also contain value labels that we may wish to search.
We can do this using the vlabs option in the lookfor_all command. The code below searches for the string "low" in the variable names, variable
labels, and value labels of all the .dta files in the

working directory.

lookfor_all low, vlabs

use "D:/Data/hsb2.dta"
then:
label list sl

use "D:/Data/hsbmis.dta"
then:
label list sl

use "D:/Data/mypars.dta"
variables: min95

Total 5 out of 5 files checked in "D:/Data/"

From this output, we can see that there are two datasets, hsb2 and hsbmis, that contain
a value label with the string "low". There's also a dataset mypars
with a variable min95 that must have "low" in its variable label.