

Parsing/interpreting Stata into SDTL

C2Metadata

Ole Voldsæter <ole.voldsater@nsd.no>

Ørnulf Risnes <ornulf.risnes@nsd.no>

Montréal 30 May 2018/IASSIST 2018

Prior experience

Development of grammar and execution of transformational and analytical commands in RAIRD / microdata.no

The screenshot shows the 'Analysemiljø - microdata.no' web application. The browser address bar displays 'https://microdata.no/rose/#'. The interface is divided into two main panels. The left panel, titled 'demo-script 4', contains a list of 16 commands for creating and manipulating a dataset named 'lansering'. The right panel shows the execution output for these commands, including a table of results for the 'regstat05' variable.

demo-script 4

```
1 create-dataset lansering
2 import BEFOLKNING_REGSTAT 2005-01-01
  as regstat05
3 tabulate regstat05
4 keep if regstat05 == '1'
5 import BEFOLKNING_KJOENN as kjønn
6 import SIVSTANDFDT_SIVSTAND
  2005-01-01 as sivst05
7 tabulate sivst05 kjønn
8 tabulate sivst05 kjønn, rowpct
9 tabulate sivst05 kjønn, freq rowpct
10 tabulate sivst05 kjønn, colpct
11 import BEFOLKNING_FOEDSELS_AAR_MND as
  faarmnd
12 generate faar = floor(faarmnd/100)
13 generate alder05 = 2005 - faar
14 summarize alder05
15 recode alder05 (min/15 = 1) (16/25 =
  2) (25/max = 3)
16 tabulate alder05
```

Execution Output:

» create-dataset lansering
Et tomt datasett, *lansering*, er opprettet og valgt

lansering» import BEFOLKNING_REGSTAT 2005-01-01 as regstat05
Importerte *regstat05* til *lansering* med 6845357 verdier

lansering» tabulate regstat05

regstat05		
1 - bosatt		4603376
3 - utvandret		405414
5 - død		1836429
9 - uregistrert person		144
Total		6845357

lansering» keep if regstat05 == '1'
2241979 enheter ble fjernet fra datasettet.

lansering» import BEFOLKNING_KJOENN as kjønn
Importerte *kjønn* til *lansering* med 4603378 verdier og 299 missingverdier

lansering» import SIVSTANDFDT_SIVSTAND 2005-01-01 as sivst05
Importerte *sivst05* til *lansering* med 4603378 verdier og 4776 missingverdier

Buttons at the bottom: >_ Send til kommandolinjen, ⚙ Kjør

The Stata language

Mostly declarative

Extremely flexible and powerful

«Convenience syntax» with abbreviations

`tab | tabu | tabul | tabula | tabulat | tabulate`

Most «state changes» created by commands can be inferred from programs/scripts alone (but not all)

Has macros (statements that generate program snippets)

Not necessarily 100% consistent grammar between different commands(?)

Handling state changes (interpreter)

DS1: V1 V2 V3 //initial «virtual» state

V1	V2	V3
2	3	0
1	4	0

>> drop V2

DS2: V1 V3 //new virtual state

V1	V3
2	0
1	0

>> generate V2 = V1 * 100

DS3: V1 V3 V2 //new virtual state

V1	V3	V2
2	0	200
1	0	100

Macros

```
replace a = b + 5 //Regular command
```

```
foreach op in "+" "-" "*" {  
    replace a = b `op' 5  
}
```

...becomes...

```
replace a = b + 5  
replace a = b - 5  
replace a = b * 5
```

The «Stata parser» is really...

- A parser
- An *interpreter* with support for on-the-fly macro precompilation
- Demo: <https://tinyurl.com/stataparser>

Biggest remaining challenges

egen

by/bysort

merge

+ unknown unknowns

Unsolvable problems

Commands that use data values as input

reshape wide

Stata 15 help for reshape

[D] reshape -- Convert data from wide to long form and vice versa

Syntax

Overview

long				wide		
+-----+				+-----+		
i j stub				i stub1 stub2		
+-----+				+-----+		
1 1 4.1			<----->	1 4.1 4.5		
1 2 4.5				2 3.3 3.0		
2 1 3.3						
2 2 3.0						
+-----+				+-----+		

Illustration from Stata 15 help