

彭文威

灵活引用:[]&

分组处理:

bysort: , ege

引用数据中的所有变量 引用某一变量的所有值 引用文件夹下所有数据

循环嵌套

while cap 让你的程序更灵活

检查

datacheck

数据清洗常用技巧

小花经济学术

彭文威

MPhil in Social Science Division of Social Science, HKUST

温州

2017 Chinese Stata Users Group Meeting 2017 年 8 月 19-20 日





目录

Stata Group Meeting

彭文威

灵活引用:[] & _n

分组处理: bysort: 、ege

引用数据中的所有变量 引用某一变量的所有值 引用文件央下所有数据

循环嵌套

cap 让你的程序更灵活

检查 assert

atacheck

- ② 分组处理: bysort: 、egen
- ③ 巧用宏与返回值 引用数据中的所有变量 引用某一变量的所有值 引用文件夹下所有数据
- 循环嵌套whilecap 让你的程序更灵活
- **⑤** 检查 assert datacheck





Stata Group Meeting 彭文威

灵活引用:[]&

引用数据中的所有变量

引用某一变量的所有值 引用文件央下所有数据 循环嵌套 cap 让你的程序更灵活 检查

分组处理:

① 灵活引用: [] & _n

② 分组处理: bysort: 、egen

⑤ 巧用宏与返回值

引用数据中的所有变量 引用某一变量的所有值 引用文件夹下所有数据

cap 让你的程序更灵活

5 检查

www.uone-tech.cn



彭文威

灵活引用:[]&

分组处理:

bysort: , ege

巧用宏与返回位

引用数据中的所有变量 引用某一变量的所有值

引用某一变量的所有值 引用文件夹下所有数据

循环嵌套

while cap 让你的程序更灵活

检查

ssert Istacheck 通过引用,自动填充。

list FamilyID ID in 10

replace FamilyID = FamilyID[_n-1] if FamilyID[_n]==.



Stata Group Meeting 彭文威

灵活引用: [] &

1 灵活引用: [] & n

⑤ 巧用宏与返回值

② 分组处理: bysort: 、egen

分组处理:

bysort: , egen

引用数据中的所有变量

引用某一变量的所有值 引用文件央下所有数据

循环嵌套

cap 让你的程序更灵活

检查

cap 让你的程序更灵活

引用数据中的所有变量 引用某一变量的所有值

引用文件夹下所有数据

检查

友万科技



分组处理: bysort: 、egen

egen income_total= max(income),by(FamilyID)

```
Stata Group
Meeting
彭文威
```

```
生成一个每个家庭人数的变量:
bysort Family: gen Num = _N
bysort Family sex: gen Num_bysex = _N
生成每个家庭的总收入:
egen Num_egen = count(ID),by(FamilyID)
egen Num_egen_bysex = count(ID),by(FamilyID sex)
生成每个家庭的总收入:
egen income_total= sum(income),by(FamilyID)
生成每个家庭的个人收入最高的收入变量:
```

```
www.uone-tech.cn
```



Stata Group Meeting

彭文威

灵活引用:[]&

_n

分组处理:

巧用宏与返回值

引用数据中的所有变量 引用某一变量的所有值 引用文件央下所有数据

循环嵌套

while cap 让你的程序更灵活

检查 assert

atacheck

- 灵活引用: [] & _n
- ② 分组处理: bysort: 、egen
- ③ 巧用宏与返回值 引用数据中的所有变量 引用某一变量的所有值 引用文件夹下所有数据
- While cap 让你的程序更灵活
- 5 检查 asser





return list

Stata Group Meeting 彭文威

灵活引用:[]&

_n

分组处理: bysort: 、egen

巧用宏与返回值

引用数据中的所有变量 引用某一变量的所有值 引用文件央下所有数据

循环嵌套

while cap 让你的程序更灵活

检查 assert return list

- creturn list
- ereturn list
- sreturn list





ds 命令

Stata Group Meeting

彭文威

灵活引用:[]&

分组处理:

bysort: , egen

· 力用宏与返回值 引用数据中的所有变量

引用某一变量的所有值 引用文件夹下所有数据

循环嵌套

while cap 让你的程序更灵活

检查

atacheck

ds // 列举出所有变量名称,并存在返回值里。ds *ID*

return list

www.uone-tech.cn



levelsof

```
Stata Group
Meeting
```

```
彭文威
```

```
灵活引用:[] &
__n
```

分组处理:

bysort: , egei

均用宏与返回值 引用数据中的所有变量

引用某一变量的所有值 引用文件夹下所有数据

循环嵌套

```
while cap 让你的程序更灵活
```

检查

```
levelsof FamilyID // 列举出变量的所有值,并存在返回值里。return list levelsof FamilyID ,local(fam) foreach family in `fam'{ .....}
```





local 扩展函数

Stata Group Meeting

彭文威

灵活引用:[]

分组处理: bysort: 、egen

巧用宏与返回值

引用数据中的所有变量 引用某一变量的所有值 **引用文件央下所有数据**

循环嵌套 while

cap 让你的程序更灵活 检查

assert datacheck In list might be returned "notes" "subpanel".
eg:
local folderslist: dir . dir "*" // . 为当前工作路径

dis `"`folderslist'"' // 列举当前工作路径下的所有文件夹名称

local list : dir . dirs "*" makes a list of all subdirectories of the current directory



local 扩展函数

Stata Group Meeting

彭文威

灵活引用:[]&

分组处理: bysort: , egen

引用数据中的所有变量 引用某一变量的所有值

引用文件央下所有数据

cap 让你的程序更灵活

检查

eg: local fileslist: dir . files "*.dta" // . 为当前工作路径 dis `"`fileslist'"' // 列举当前工作路径下的所有 dta 文件名称

> www.uone-tech.cn 友万科技

local list : dir .files "*" makes a list of all regular files in the current directory.

彭文威

```
灵活引用:[]&
```

```
分组处理:
```

引用数据中的所有变量

引用某一变量的所有值 引用文件来下所有数据

检查

循环嵌套

cap 让你的程序更灵活

对当前路径下所有子文件夹的所有 dta 文件进行循环操作。

```
foreach folder in `folderslist' {
local fileslist: dir `".\`folder'"' files "*.dta"
foreach file in `fileslist' {
use `file',clear
. . . . .
```

local folderslist: dir . dir "*"



盾环

Stata Group Meeting

彭文威

```
_n
```

分组处理: bysort: 、egen

巧用宏与返回值 引用数据中的所有变量 引用某一变量的所有值

引用文件央下所有数据 循环嵌套

while cap 让你的程序更灵活

检查 assert

```
对当前路径下所有子文件夹的所有 dta 文件进行循环操作。
```

```
local folderslist: dir . dir "*"
foreach folder in `folderslist' {

local fileslist: dir `".\`folder'"' files "*.dta"
foreach file in `fileslist' {
  use `file',clear
  ds
foreach var in `r(varlist)'{
   .....
}

}
}
```





- Stata Group Meeting 彭文威
- 灵活引用: [] &
- 分组处理:

- 引用数据中的所有变量 引用某一变量的所有值
- 引用文件央下所有数据 循环嵌套
- cap 让你的程序更灵活
- 检查

- 1 灵活引用: [] & n
- ② 分组处理: bysort: 、egen
 - 3 巧用宏与返回值 引用数据中的所有变量 引用某一变量的所有值 引用文件夹下所有数据
 - 4 循环嵌套 while
 - cap 让你的程序更灵活
 - 5 检查





Stata Group Meeting 彭文威

灵活引用: [] &

分组处理:

引用数据中的所有变量

引用某一变量的所有值

引用文件央下所有数据 循环嵌套

cap 让你的程序更灵活 检查

1 灵活引用: [] & n

② 分组处理: bysort: 、egen

3 巧用宏与返回值 引用数据中的所有变量 引用某一变量的所有值 引用文件夹下所有数据

cap 让你的程序更灵活

6 检查 assert

datacheck

www.uone-tech.cn 友万科技



assert

Stata Group Meeting

彭文威

灵活引用: []

分组处理:

巧用宏与返回值

引用数据中的所有变量 引用某一变量的所有值 引用文件来下所有数据

循环嵌套

cap 让你的程序更灵活

检查

atacheck

假设,每个家庭的成员 ID 是以年龄大小顺序排列的,比如,FamilyID 为 1 的家庭有 4 个成员,成员 ID 分别为 1、2、3、4。则成员的年龄排序为 1>2>3>4。当出现不连续的排序时,为数据录入错误,则要舍弃。数据量大,肉眼识别易错且效率低下。

bysort FamilyID: gen index= _n
assert index==ID
return list

assert 命令将对其后的表达式进行判断,并返回多少个判断错误。





datacheck

Stata Group Meeting 彭文威

分组处理:

引用某一变量的所有值

循环嵌套

cap 让你的程序更显活

检查

datacheck

同 assert 命令, 但功能更强大。假设, 每个家庭的成员 ID 是以年龄大小顺序排 列的, 比如, FamilyID 为 1 的家庭有 4 个成员, 成员 ID 分别为 1、2、3、4。 则成员的年龄排序为 1>2>3>4。当出现不连续的排序时,为数据录入错误,则 要舍弃。

bysort FamilyID: gen index= _n assert index==ID return list

datacheck 命令将对其后的表达式进行判断, 返回多少个判断错误并列举。





彭文威

灵活引用:[]&

分组处理: bysort: 、egen

bysort: , egen

引用数据中的所有变量 引用某一变量的所有值 引用文件夹下所有数据

循环嵌套

while

cap 让你的程序更灵活

检查

datacheck

Thank you!

