厦門大學 XIAMEN UNIVERSITY
厦门大学 管|理|学|院 SCHOOL OF MANAGEMENT, XIAMEN UNIVERSITY
中国能源政策研究院 CHINA INSTITUTE FOR STUDIES IN ENERGY POLICY

framerge：通过数据框进行数据横向合并

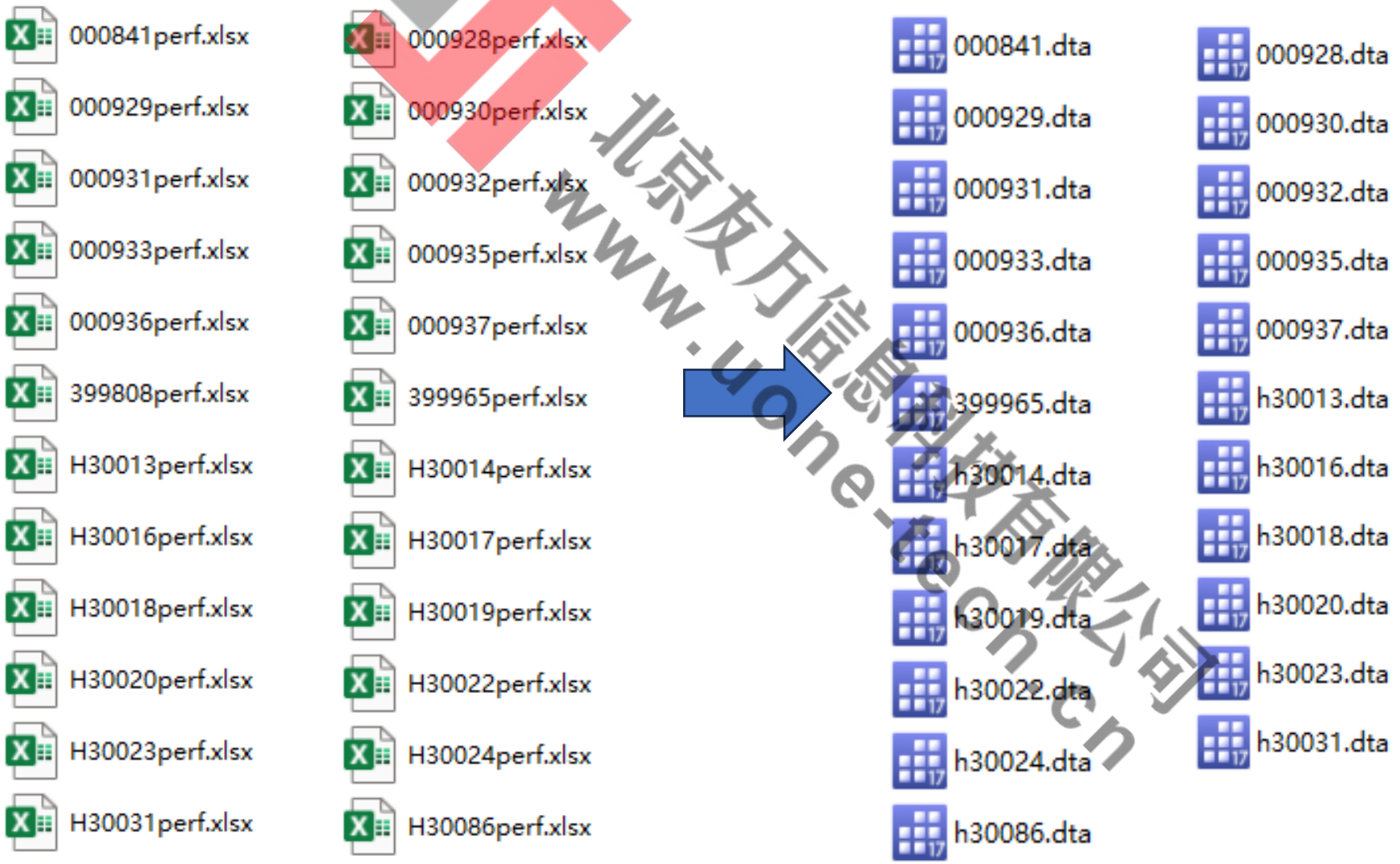
杜克锐　陈巧雯
中国能源政策研究院

第八届Stata中国用户大会
2024年8月20日

为什么要做这个命令?

- merge?
- joinby?
- frlink + frget?

merge? joinby?



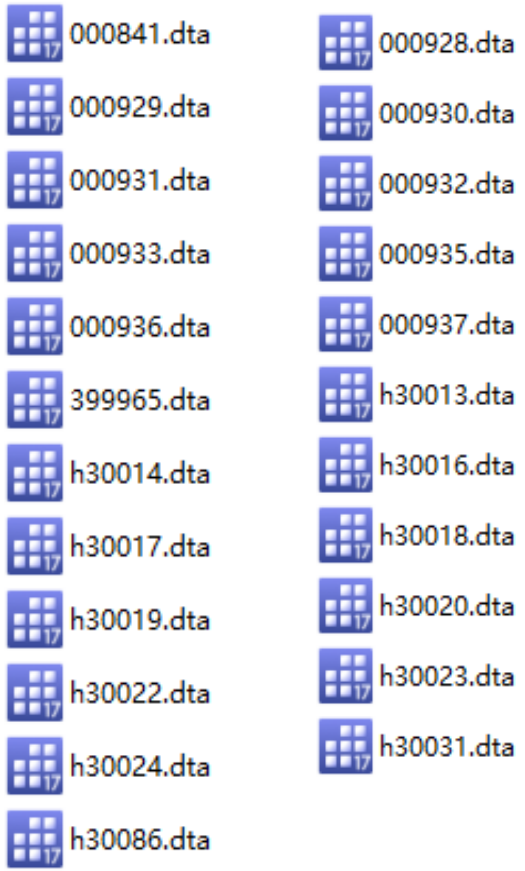
- 本地保存
- 中间文件

frlink + frget?


```
frlink {1:1|m:1} varlist1, frame(frame2 [varlist2])
                 [generate(linkvar1)]
```


1:m
m:m

→ 本地保存 → merge
joinby →

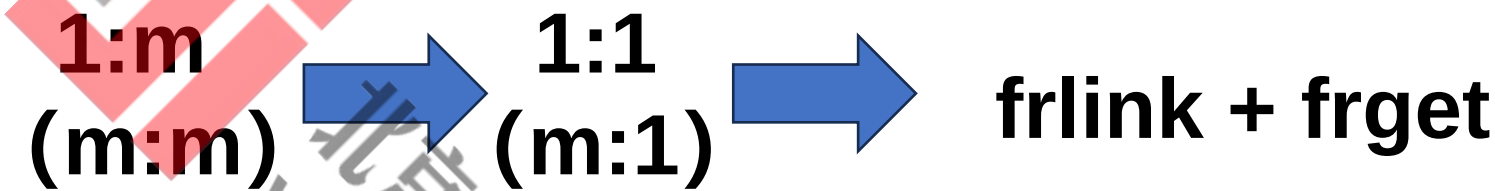


为什么要做这个命令?

- merge & joinby：本地保存，时间及操作成本、中间文件
- frlink + frget：无法处理 1:m 和 m:m 关系
- 工作效率

- 目的:
  - 通过frame在内存中直接操作数据，无需读写硬盘
  - 多种数据合并关系：1:1，m:1，1:m和 m:m

总体思路



- 生成一个行号变量来唯一标识使用数据框中的观测值
- 将主数据框的观测值扩展到相应记录的数量
- 用相应的行号值识别主数据框中的观测值
- 以行号变量为标识变量执行frlink和frget

思路：1:m

Step 1: 在使用数据框生成行号变量

```
. frame discharge2{
. sort patientid
. gen rownum = _n
. frame put patientid rownum, into(tempdata)
.                         }
```

思路：1:m

Step 2 记录每个初始id第一行和最后一行行号

```
. frame tempdata {
. bys patientid (rownum): gen row1 = _n
. bys patientid (rownum): gen row2 = _N
. keep if row1==row2
(7,920 observations deleted)
. replace row1 = rownum -row2+1
(1,980 real changes made)
. replace row2 = rownum
(1,979 real changes made)
. drop  rownum
.                       }
```

Step 3 把第一行和最后一行行号放到主数据框

```
. webuse discharge1, clear
(Fictional WA hospital discharges)

. replace patientid=patientid+100000 in 1
(1 real change made)

. frlink 1:1 patientid, frame(tempdata)
(1 observation in frame default unmatched)

. frget row1 row2, from(tempdata)
(1 missing value generated)
(1 missing value generated)
(2 variables copied from linked frame)
```

Step 4 主数据框：数据扩展，并计算扩展后每个观测记录行号变量对应值

```
. gen nreps = row2-row1+1
(1 missing value generated)

. expand nreps
(1 missing count ignored; observation not deleted)
(7,916 observations created)

. bys patientid: gen rownum = row1 +_n-1
(1 missing value generated)

. drop row1 row2 nreps
```

思路：1:m

Step 5: 连接主数据框和使用数据框

```
. frlink 1:1 rownum, frame(discharge2)
(1 observation in frame default unmatched)
```

Step 6: 从使用数据框获取变量

```
. frget date, from(discharge2)
(1 missing value generated)
(1 variable copied from linked frame)
```

思路：m:m

- 主数据框扩展后观测记录对应行号变量值计算
  - 主数据框初始id无法唯一标识观测记录
  - 初始id + 相同id组内顺序
- 其余和1:m相同

```
frame reset
cap frame create discharge2
frame discharge2{
    webuse discharge2,clear
    expand 5
    bys patientid: gen date =_n
    gen rownum = _n
    frame put patientid rownum, into(tempdata)
}

frame tempdata {
    bys patientid (rownum): gen row1 = _n
    bys patientid (rownum): gen row2 = _N
    keep if row1==row2
    replace row1 = rownum -row2+1
    replace row2 = rownum
    drop  rownum

}
```

```
webuse discharge1, clear
expand 3
replace patientid=patientid+100000 in 1
frlink m:1 patientid, frame(tempdata)
frget row1 row2, from(tempdata)

gen nreps = row2-row1+1
bys  patientid: gen gj = _n
expand nreps
bys patientid gj: gen rownum = row1 +_n-1
drop row1 row2 nreps

frlink m:1 rownum, frame(discharge2)
frget date, from(discharge2)

sort patientid date
list patientid age date in 1/25
list patientid age date in 29696
```

framerge 1:m

discharge1

```
. list patientid age in 1/3
```

	patien~d	age
1.	100001	77
2.	2	81
3.	3	37

discharge2

```
. list patientid date in 1/15
```

	patien~d	date
1.	1	1
2.	1	2
3.	1	3
4.	1	4
5.	1	5
6.	2	1
7.	2	2
8.	2	3
9.	2	4
10.	2	5
11.	3	1
12.	3	2
13.	3	3
14.	3	4
15.	3	5

```
. framerge 1:m patientid, frame(discharge2) get(date)
. list patientid age date in 1/10
```

	patien~d	age	date
1.	2	81	1
2.	2	81	2
3.	2	81	3
4.	2	81	4
5.	2	81	5
6.	3	37	1
7.	3	37	2
8.	3	37	3
9.	3	37	4
10.	3	37	5

```
. list patientid age date in 9896
```

	patien~d	age	date
9896.	100001	77	.

framerge m:m

厦門大學 XIAMEN UNIVERSITY

discharge1

```
. list patientid age in 1/5
```

	patien~d	age
1.	1	77
2.	1	77
3.	2	81
4.	2	81
5.	2	81

```
. list patientid age in 5940
```

	patien~d	age
5940.	100001	77

discharge2

```
. list patientid date in 1/15
```

	patien~d	date
1.	1	1
2.	1	2
3.	1	3
4.	1	4
5.	1	5
6.	2	1
7.	2	2
8.	2	3
9.	2	4
10.	2	5
11.	3	1
12.	3	2
13.	3	3
14.	3	4
15.	3	5

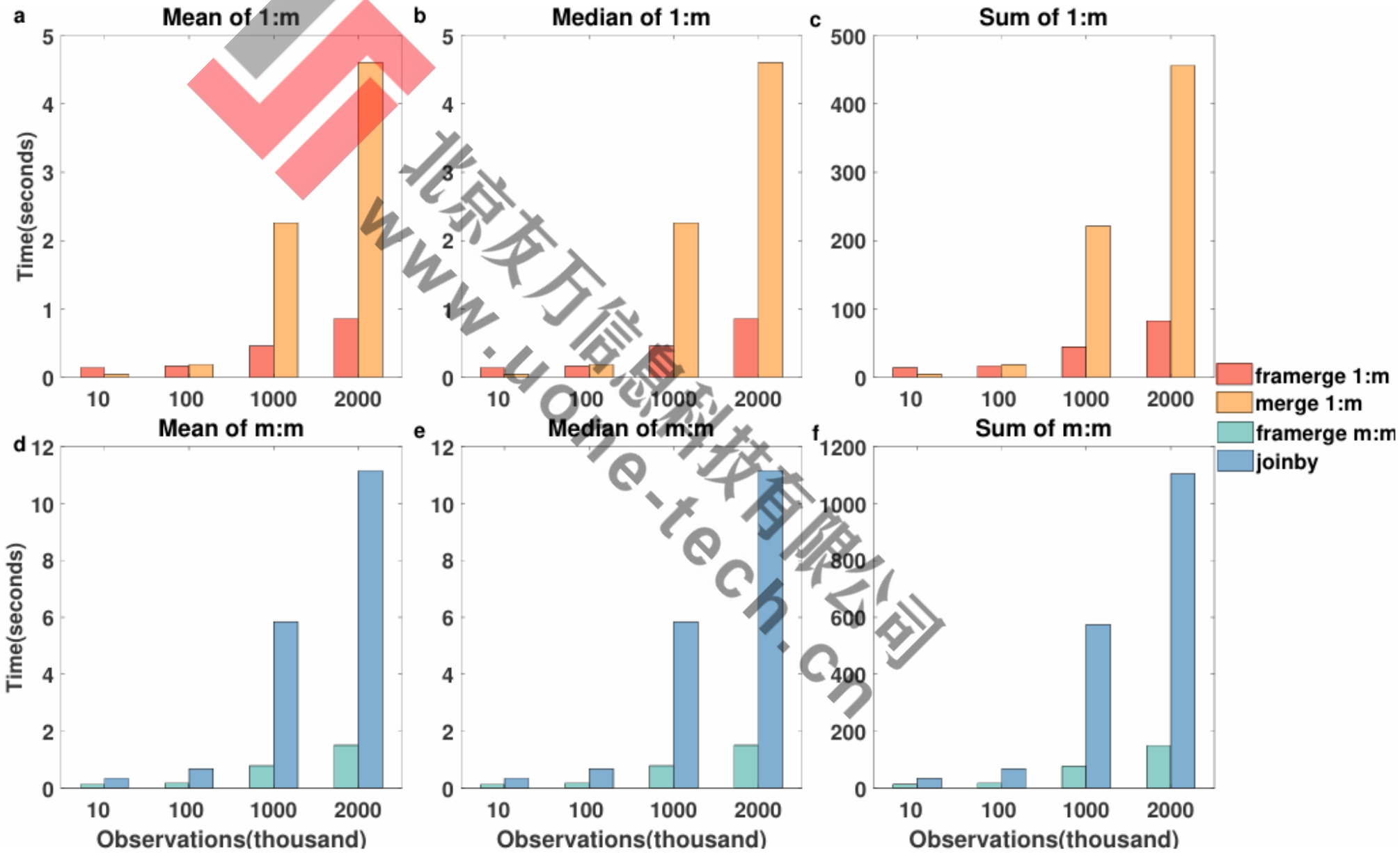
```
. framerge m:m patientid, frame(discharge2) get(date)
. list patientid age date in 1/25
```

	patien~d	age	date
1.	1	77	1
2.	1	77	1
3.	1	77	2
4.	1	77	2
5.	1	77	3
6.	1	77	3
7.	1	77	4
8.	1	77	4
9.	1	77	5
10.	1	77	5
11.	2	81	1
12.	2	81	1
13.	2	81	1
14.	2	81	2
15.	2	81	2
16.	2	81	2
17.	2	81	3
18.	2	81	3
19.	2	81	3
20.	2	81	4
21.	2	81	4
22.	2	81	4
23.	2	81	5
24.	2	81	5
25.	2	81	5

```
. list patientid age date in 29696
```

	patien~d	age	date
29696.	100001	77	.

framerge处理大数据集




那么...在哪里
才能买得到呢?

framerge安装

- **安装代码**

net install kgitee, from(https://gitee.com/kerrydu/kgitee/raw/master) replace

kgitee framerge

- **Gitee主页**

https://gitee.com/kerrydu/kgitee/tree/master/

感谢批评指正

杜克锐
kerrydu@xmu.edu.cn

陈巧雯
chenqiaowen@stu.xmu.edu.cn