A `diff` Command for Use with Data Files

L. Philip Schumm Michael S. Johnson

Department of Health Studies
University of Chicago

July 24, 2006

Example: Auto data

	make	~~price~~	mpg	rep78	obs
1.	AMC Concord	~~4,099~~	22->24	3	1
2.	AMC Pacer	~~4,749~~	17	3	2
3.	AMC Spirit	~~3,799~~	22	.->2	3
4.	Buick Century	~~4,816~~	20->.	3	4
~~5.~~	~~Buick Electra~~	~~7,827~~	~~15~~	~~4~~	
etc.					
75.	Volvo 240DL Wagon		18	5	75

New command: updatedata

```
. sysuse auto
(1978 Automobile Data)

. updatedata (make) using modified_auto, detail

master: 74 obs, using: 74 obs, matched: 73 obs
+ 1 obs
- 1 obs

price: dropped
mpg: 1 changes, 0 miss -> nonmiss, 1 nonmiss -> miss
     obs 1 (make = "AMC Concord") 22 -> 24
     obs 4 (make = "Buick Century") 20 -> .
rep78: 0 changes, 1 miss -> nonmiss, 0 nonmiss -> miss
       obs 3 (make = "AMC Spirit") . -> 2
headroom: identical
trunk: identical
...

obs: added
```

updatedata (cont.)

- Similar to `PROC COMPARE` in SAS

updatedata (cont.)

- Similar to `PROC COMPARE` in SAS
- Useful in several contexts
  - multiple spreadsheets

updatedata (cont.)

- Similar to `PROC COMPARE` in SAS
- Useful in several contexts
  - multiple spreadsheets
  - double data entry (verification)

Application to large-scale data management

Example: NIH-funded IBD Genetics Consortium

Application to large-scale data management

Example: NIH-funded IBD Genetics Consortium

- Several Genetic Research Centers (GRCs)

Application to large-scale data management

Example: NIH-funded IBD Genetics Consortium

- Several Genetic Research Centers (GRCs)
- Each accrues subjects, collects blood and phenotype data

Application to large-scale data management

Example: NIH-funded IBD Genetics Consortium

- Several Genetic Research Centers (GRCs)
- Each accrues subjects, collects blood and phenotype data
- Data Coordinating Center (DCC), responsible for:

Application to large-scale data management

Example: NIH-funded IBD Genetics Consortium

- Several Genetic Research Centers (GRCs)
- Each accrues subjects, collects blood and phenotype data
- Data Coordinating Center (DCC), responsible for:
  - Collecting and integrating data from GRCs

Application to large-scale data management

Example: NIH-funded IBD Genetics Consortium

- Several Genetic Research Centers (GRCs)
- Each accrues subjects, collects blood and phenotype data
- Data Coordinating Center (DCC), responsible for:
  - Collecting and integrating data from GRCs
  - Maintaining central, up-to-date database

Application to large-scale data management

Two extensions:

Application to large-scale data management

Two extensions:

- Multiple identifiers

Application to large-scale data management

Two extensions:

- Multiple identifiers
  - pedigree ID and individual ID

Application to large-scale data management

Two extensions:

- Multiple identifiers
  - pedigree ID and individual ID
  - local sample ID

Application to large-scale data management

Two extensions:

- Multiple identifiers
  - pedigree ID and individual ID
  - local sample ID
  - repository "K" number

Application to large-scale data management

Two extensions:

- Multiple identifiers
  - pedigree ID and individual ID
  - local sample ID
  - repository "K" number
  - central Consortium ID

Problems with multiple identifiers

master

	id1	id2
1.	1	a
2.	2	b
3.	3	c
4.		c
5.	4	d
6.		e

using

	id1	id2
1.	1	a
2.		b
3.	2	
4.	3	c
5.	4	f
6.	5	
7.	6	g
8.	6	

Problems with multiple identifiers

master

	id1	id2
1.	1	a
2.	2	b
3.	3	c
4.		c
5.	4	d
6.		e

using

	id1	id2
1.	1	a
2.		b
3.	2	
4.	3	c
5.	4	f
6.	5	
7.	6	g
8.	6	

- one-to-many

Problems with multiple identifiers

master

	id1	id2
1.	1	a
2.	2	b
3.	3	c
4.		c
5.	4	d
6.		e

using

	id1	id2
1.	1	a
2.		b
3.	2	
4.	3	c
5.	4	f
6.	5	
7.	6	g
8.	6	

- one-to-many
- many-to-one

Problems with multiple identifiers

master

	id1	id2
1.	1	a
2.	2	b
3.	3	c
4.		c
5.	4	d
6.		e

using

	id1	id2
1.	1	a
2.		b
3.	2	
4.	3	c
5.	4	f
6.	5	
7.	6	g
8.	6	

- one-to-many
- many-to-one
- discrepancy

Problems with multiple identifiers

master

	id1	id2
1.	1	a
2.	2	b
3.	3	c
4.		c
5.	4	d
6.		e

using

	id1	id2
1.	1	a
2.		b
3.	2	
4.	3	c
5.	4	f
6.	5	
7.	6	g
8.	6	

- one-to-many
- many-to-one
- discrepancy
- duplicate

Problems with multiple identifiers

master

	id1	id2
1.	1	a
2.	2	b
3.	3	c
4.		c
5.	4	d
6.		e

using

	id1	id2
1.	1	a
2.		b
3.	2	
4.	3	c
5.	4	f
6.	5	
7.	6	g
8.	6	

- one-to-many
- many-to-one
- discrepancy
- duplicate
- possible duplicate

Application to large-scale data management (cont.)

Two extensions:

- Multiple identifiers

Application to large-scale data management (cont.)

Two extensions:

- Multiple identifiers
- do-file "patch" (automatically generated)

Application to large-scale data management (cont.)

Two extensions:

- Multiple identifiers
- do-file "patch" (automatically generated)
  - apply and track changes

Application to large-scale data management (cont.)

Two extensions:

- Multiple identifiers
- do-file "patch" (automatically generated)
  - apply and track changes
  - receipt

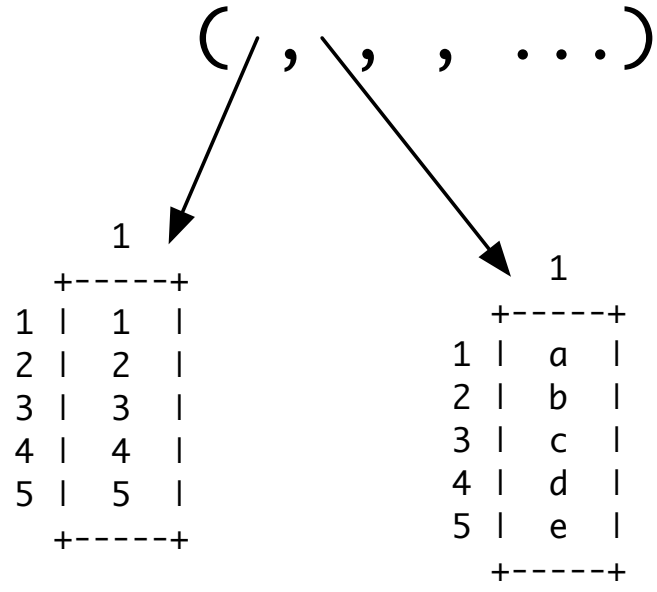
Full syntax

```
updatedata (id1) (id2) (...) [varlist] using <filename>,
    [dofile(<dofilename>) addobs dropobs addvars dropvars
    set2miss detail]
```

Storing dataset containing numeric and string vars

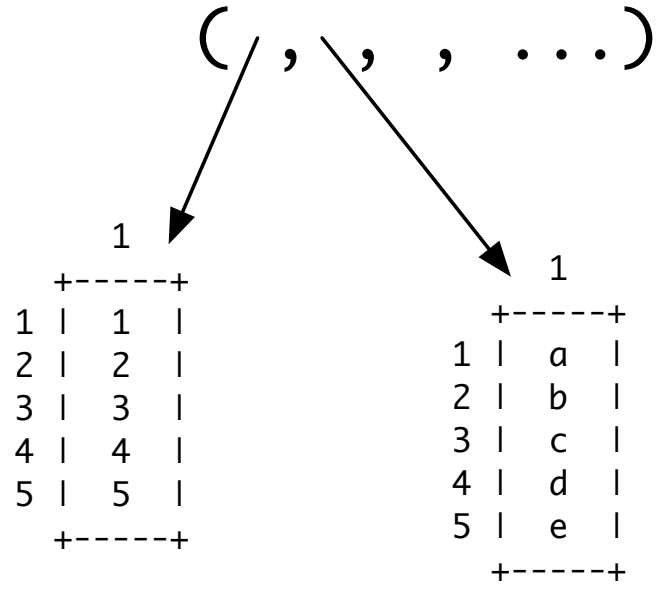
Storing dataset containing numeric and string vars

- rowvector of pointers to colvectors



Storing dataset containing numeric and string vars

- rowvector of pointers to colvectors



- dataset structure

```
struct ud_dataset {
    string rowvector                    varnames
    pointer(colvector) rowvector        data
}
```

Use of new `datasignature` command

Use in two ways:

Use of new `datasignature` command

Use in two ways:

- preliminary check for identical data (fast)

Use of new `datasignature` command

Use in two ways:

- preliminary check for identical data (fast)
- construct check at top of patch file (i.e., do-file)

Limitations of `cf`

Limitations of `cf`

- files must have same number of observations

Limitations of `cf`

- files must have same number of observations
- files must be in same sort order

Limitations of `cf`

- files must have same number of observations
- files must be in same sort order
- ignores variables not in master

Limitations of `cf`

- files must have same number of observations
- files must be in same sort order
- ignores variables not in master
- summary output only

Modifications to `updatedata`

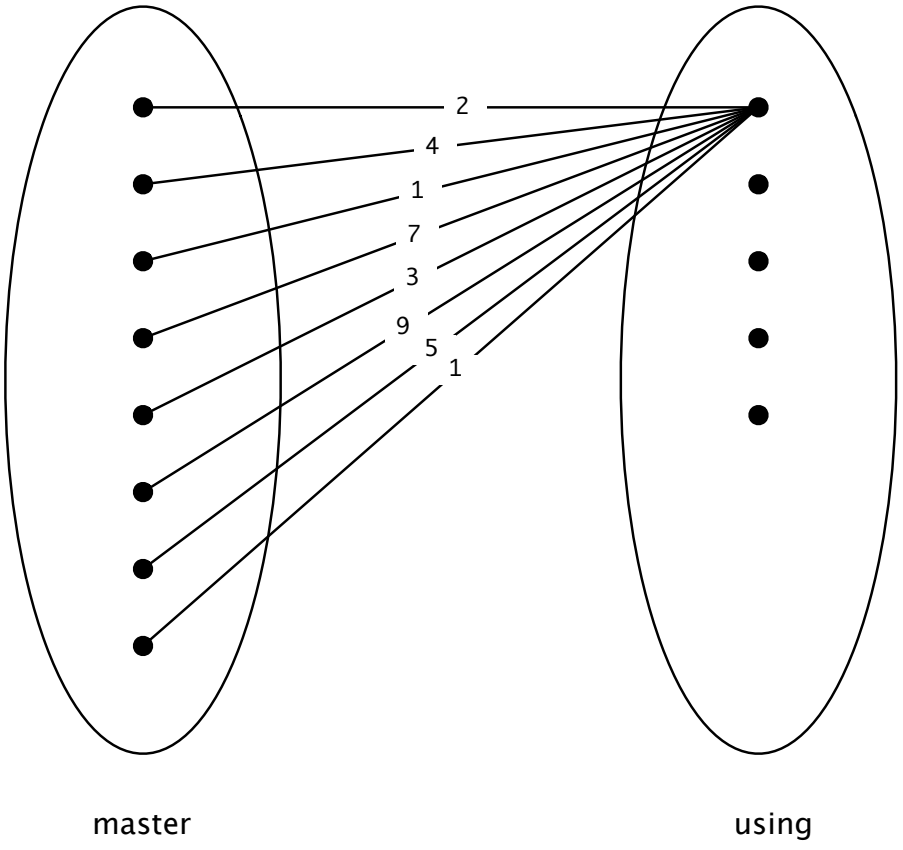
Modifications to `updatedata`

- Construct pseudo-identifier consisting of all common vars together with `_n` (within by-group)

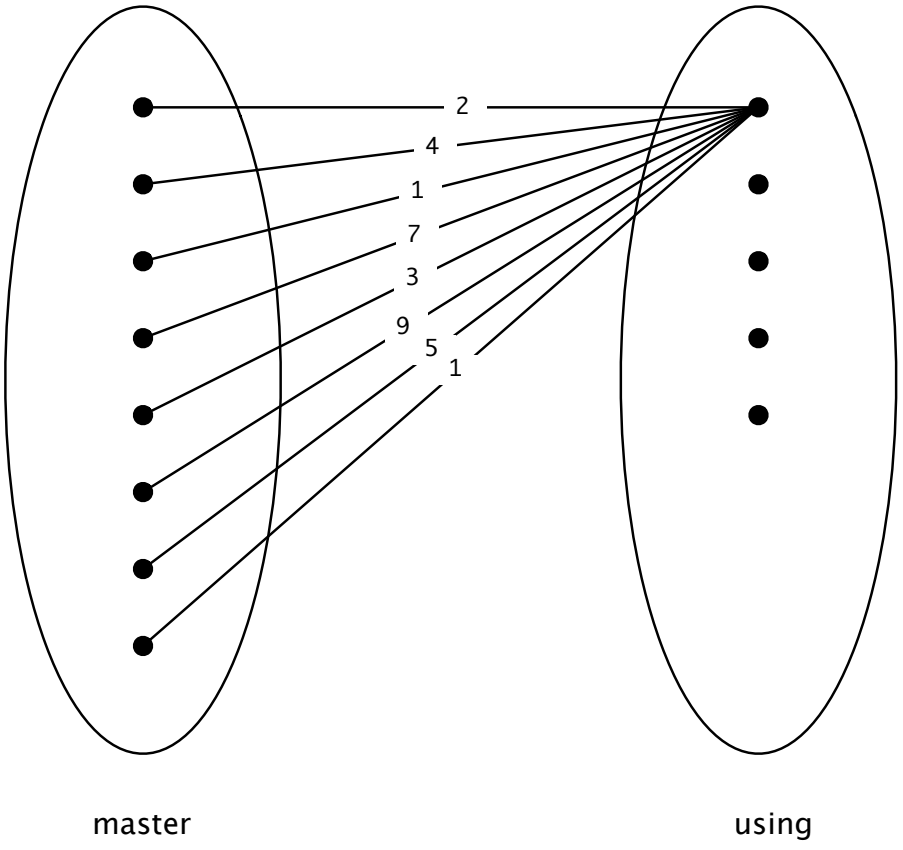
Modifications to `updatedata`

- Construct pseudo-identifier consisting of all common vars together with `_n` (within by-group)
- After all identical observations matched, match *all* remaining obs according to "optimal" criterion

Weighted bipartite graph



Weighted bipartite graph



- choose matching to minimize sum of edit distances

datadiff

Syntax:

```
datadiff [varlist] using <filename>
```