

## Data Management in Stata

- Keep all of the data and files for each project in the project's own folder.
- When making corrections or other changes to the original ('master') data set, document/log the corrections/changes in a do-file and save the corrections/changes under a *new* name (i.e. as a *new* data set, along with a backup) that is clearly identifiable as the successor to the master data set.
  - *Always leave the master data in its original condition.*
  - As you continue to make changes to the data, continue to document/log the changes in successive do-files/log-files, and *continue to save the changes under new data-set names whose sequence is clearly identifiable and clearly linked to each successive do-file:*
    - E.g., master.dta; data1 = do1 (log do1); data2 = do2 (log do2); data3 = do3 (log do3).
    - Erase unnecessary data sets at the end of each do-file; you can later reproduce them if necessary by invoking 'do master'.

Let's say that your project will analyze data3. *Ideally you'll manage the data by erasing data1-data2 at the end of each working session. You can later reproduce them if necessary by invoking 'do master'.*

### Example

---

\* master.do. Uses master.dta to create data1, data2, & data3. master.dta = ///  
hsb2.dta, November 2007.

```
capture log close  
set more off
```

```
do do1
```

```
do do2
```

```
do do3
```

---

\* do1.do. Uses master.dta to create data1.dta. master.dta=hsb2.dta. ///  
November 2007.

```
version 10.0  
capture log close  
set more off  
log using do1, replace  
use master, clear
```

```
su  
gen ms=math+science  
su math science ms  
gr box ms  
gr save grboxms, replace  
hist ms, norm,
```

```
gr save grhistms, replace

la var ms "math + science"
note ms: TS "gen ms=math+science"
placevar ms, a(science)

su
compress
save data1, replace

log close
exit
```

---

\* do2.do. Uses data1 to create data2. data1 is created in do1. November 2007.

```
version 10.0
capture log close
set more off
log using do2, replace
use data1, clear
```

```
su
gen rw=read+write
su read write rw
gr box rw
sa grboxrw, replace
histogram rw, norm
sa grhistrw, replace
```

```
la var rw "read + write"
note rw: TS gen rw=read+write"
placevar rw, a(write)
```

```
su
compress
save data2, replace
erase data1.dta
```

```
log close
exit
```

---

\* do3.do. Uses data2 to create data3. data2 is created in do2. November 2007.

```
version 10.0
set more off
log using do3, replace
use data2, clear
```

```
su
gen rwms=read+write+math+science
su read write math science rwms
```

```
gr box rwms
sa grboxrwms, replace
hist rwms, norm
sa grhistrwms, replace

la var rwms "read+write+math+science"
note rwms: TS gen rwms = read + write + math + science"
placevar rwms, a(science)

su
compress
save data3, replace
erase data2.dta

log close
exit
```