

WORKFLOW

Effective Data Management Strategies for Doing Research Well



WHY?

Why an explicit focus on workflow?

WHAT?

What are the steps and tasks in an effective workflow?

HOW?

How can we use Stata and other programs to accomplish these steps?

Who Am I?

Eldest of four...



Wife...



Mother of three...



I've been bossin' people around since 1978...

My Experience = My Institutional Knowledge

- BA Psychology & Sociology – BGSU, 2001
- MA Social Psychology – BGSU, 2006
- Dissertation Fellow – NCFMR, 2008/2009
- Research Assistant – NCFMR, 2009/2010
- Social Science Data Analyst – NCFMR, 2010-Present
- PhD Family Sociology – BGSU, 2012

BUT...

I too, am still
learning.

Discipline

Self-Control
Self-Restraint

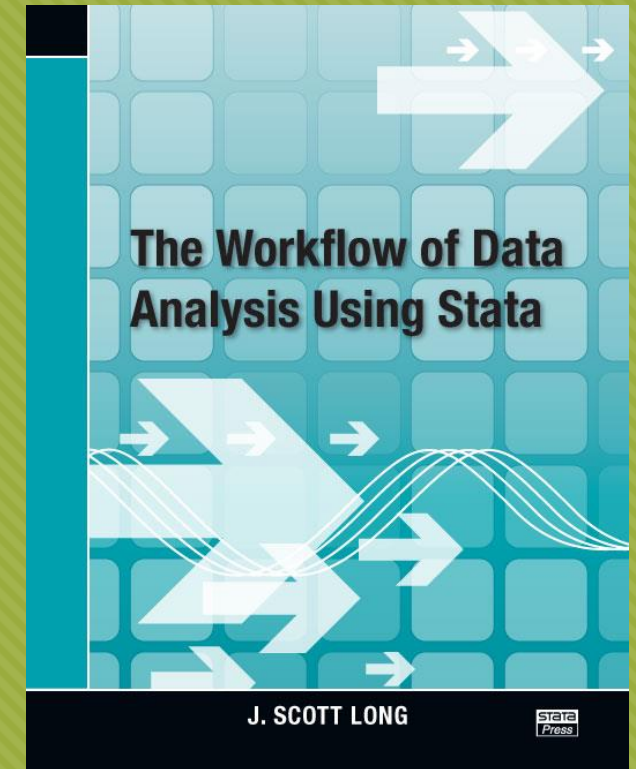
Define: WORKFLOW

- **WORKFLOW** is the series of activities that are necessary to complete a task.

Tasks

- Class papers
- Class assignments
- Journal Articles
- Conference Posters
- Presentations
- Family Profiles
- Ohio Population News
- Outside Requests for Information
- Inside Requests for Information
- Questionnaire/Survey Development
- Data Collection

The Workflow of Data Analysis Using Stata



J. Scott Long

WHY?

Why an explicit focus on workflow?

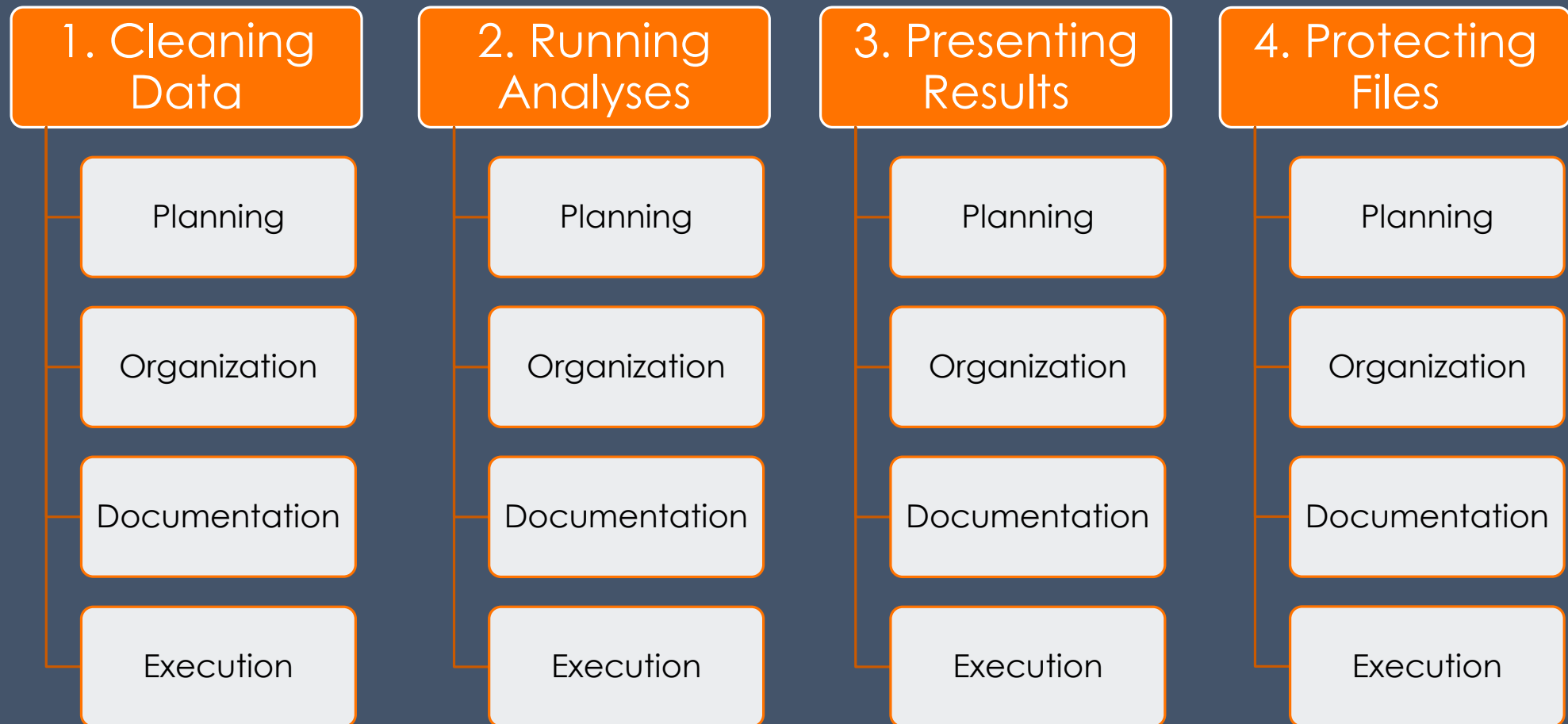
Replication

- The Guiding Principle for Workflow
 - Cornerstone
 - Science demands replicability

WHAT?

What are the steps and tasks in an effective workflow?

What? What are the steps and tasks in an effective workflow?



HOW?

How can we use Stata and other programs to accomplish these steps?

HOW?

How can we use Stata and other programs to accomplish these steps?

- Recall the four tasks:

1. Planning
2. Organization
3. Documentation
4. Execution

HOW?

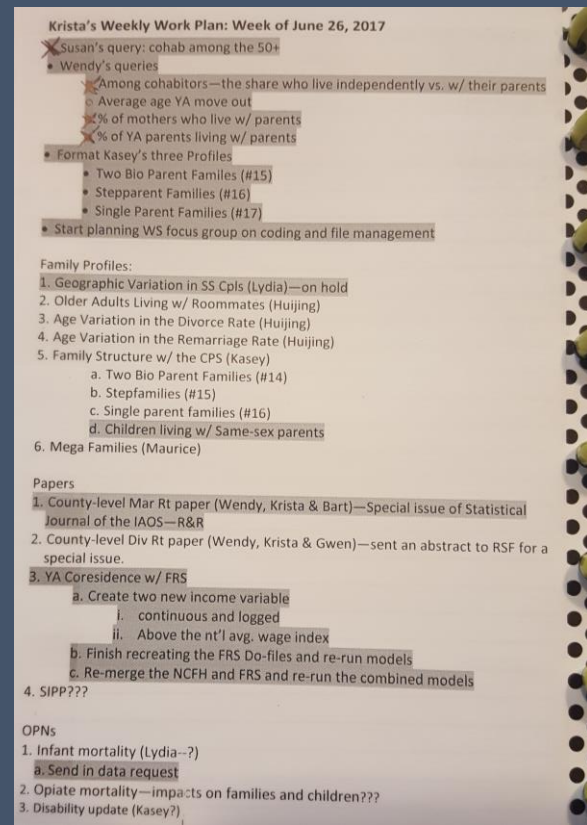
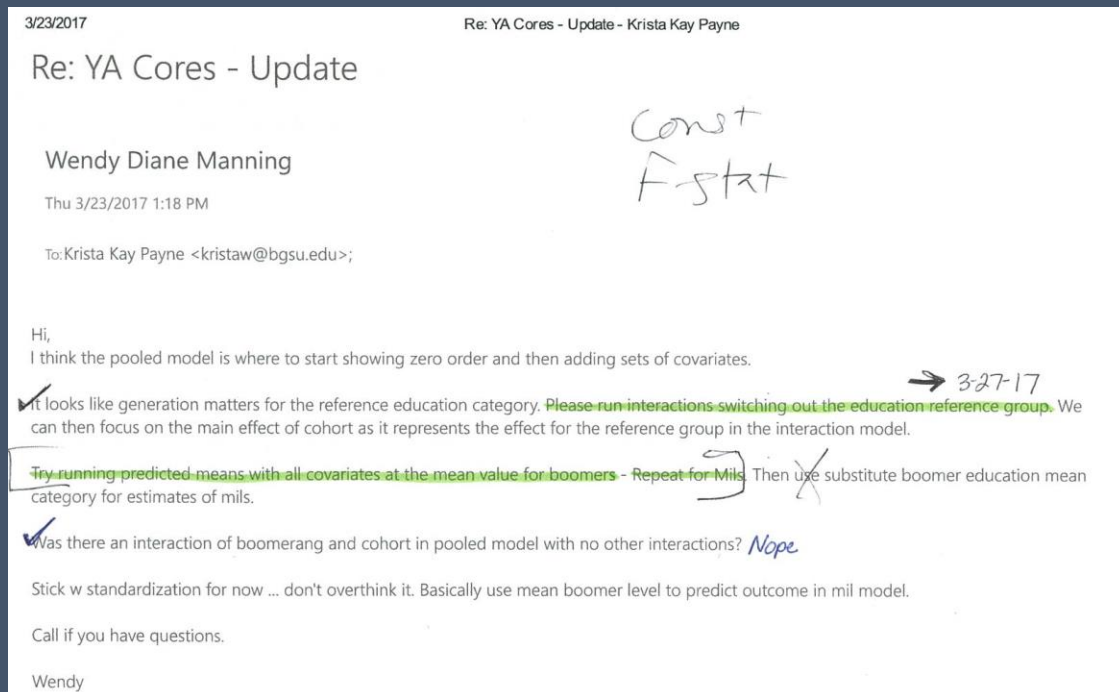
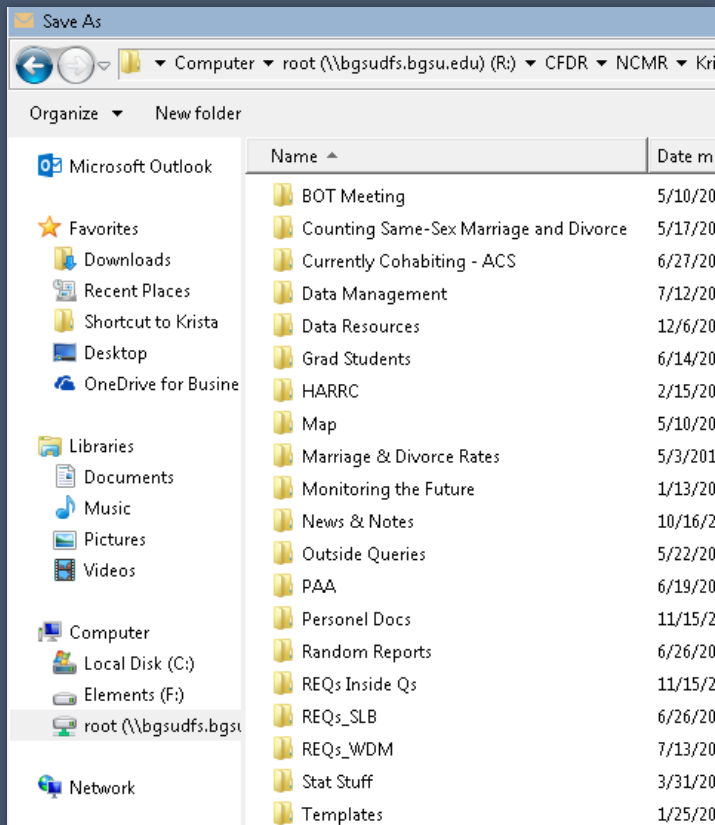
How can we use Stata and other programs to accomplish these steps?

1. Planning

- What types of analyses are needed?
- How will you handle missing data?
- What new variables must be constructed?

HOW? How can we use Stata and other programs to accomplish these steps?

Planning



HOW?

How can we use Stata and other programs to accomplish these steps?

Planning

Krista's Weekly Work Plan: Week of June 26, 2017

- Susan's query: cohab among the 50+
- Wendy's queries
 - Among cohabitators—the share who live independently vs. w/ their parents
 - Average age YA move out
 - % of mothers who live w/ parents
 - % of YA parents living w/ parents
- Format Kasey's three Profiles
 - Two Bio Parent Families (#15)
 - Stepparent Families (#16)
 - Single Parent Families (#17)
- Start planning WS focus group on coding and file management

HOW?

How can we use Stata and other programs to accomplish these steps?

2. Organization

- Requires you to think systematically about:
 - How you name files and variables
 - How you organize directories
 - How you keep track of which computer has what information
 - Where you store research materials

HOW?

How can we use Stata and other programs to accomplish these steps?

Organization: Naming Files

- Only use
 - LETTERS
 - NUMBERS
 - UNDERSCORES
 - DASHES
 - SPACES

Keyword Title Initials

FP_HMI_01-16-14_kkp

Suffix MM-DD-YY

The diagram illustrates the components of the file name 'FP_HMI_01-16-14_kkp'. Brackets above the text identify 'FP' as the 'Suffix', 'HMI' as the 'Keyword', and '01-16-14' as the 'Title' (date in MM-DD-YY format). A bracket below the text identifies 'kkp' as the 'Initials'.

HOW?

How can we use Stata and other programs to accomplish these steps?

Organization: Organizing files and folders (i.e., directories)

- Sound familiar?
 - You have multiple versions of a file and you do not know which is which.
 - You cannot find a file and you think you may have deleted.
 - You and a colleague are not sure which draft of your paper is the latest or find that there are two different “latest” or “final” drafts.

HOW?

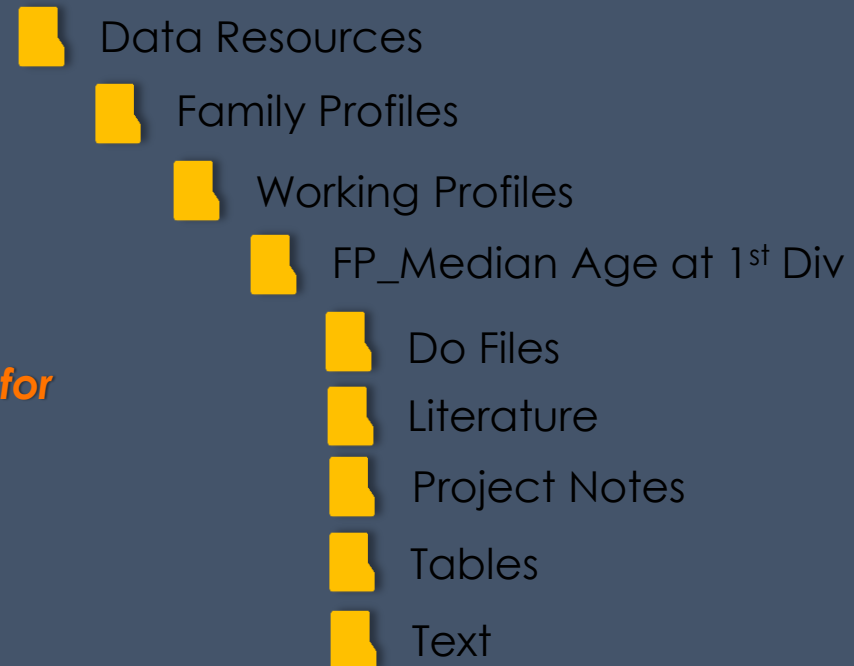
How can we use Stata and other programs to accomplish these steps?

Organization: Organizing files and folders (i.e., directories)

- Solution
 1. Start with a carefully designed folder structure.
 2. When files are created, place them in the correct folder.



A clear directory structure is particularly important for collaborative projects where things can get disorganized quickly.



HOW?

How can we use Stata and other programs to accomplish these steps?

Organization: Organizing files and folders (i.e., directories)



Pick a **Mnemonic** for Each Project

Examples:

- MBUD – Manning & Brown Update
- YACores – Young Adult Coresidence
- RQ – Relationship Quality
- CLMar – County-level Marriage Rates
- HMI – Health Marriage Initiative

HOW?

How can we use Stata and other programs to accomplish these steps?

Documentation: Keeping track of what you have done and thought

“ “ *Long's law of documentation: It is always faster to document it today than tomorrow.*

HOW?

How can we use Stata and other programs to accomplish these steps?

Documentation: Keeping track of what you have done and thought

- What should you document?
 - Data sources
 - Data decisions
 - Statistical analysis
 - Software
 - Storage

HOW?

How can we use Stata and other programs to accomplish these steps?

Documentation: Keeping track of what you have done and thought

- Levels of documentation?
 - The research log
 - Codebooks
 - Dataset documentation
 - Documenting do-files
 - Internally labeling documents

HOW?

How can we use Stata and other programs to accomplish these steps?

Documentation: Keeping track of what you have done and thought

- Suggestions for writing documentation?
 - Do it TODAY
 - Check it later
 - Know where the documentation is
 - Include full dates and names

HOW?

How can we use Stata and other programs to accomplish these steps?

Documentation: Keeping track of what you have done and thought

Example: Research Log

First complete set of analysis for FLIM measures paper

f2alt01a.do - 24May2002

Descriptive information on all rhs, lhs, and flim measures

f2alt01b.do - 25May2002

Compute bic' for each of four outcomes and all flim measures.

```
** Outcome: Can Work                global lhs "qcanwrk95"
** Outcome: Work in three categories global lhs "dhlthwk95"
** Outcome: bath trouble             global lhs "bathdif95"
** Outcome: adlsum95 - sum of adls   global lhs "adlsum95"
```

f2alt01c.do - 25May2002

Compute bic' for each of four outcomes and with only these restricted flim measures.

```
* 1. ln(x+.5) and ln(x+1)
* 2. 9 counts: >=5=5 >=7=7 (50% and 75%)
* 3. 8 counts: >=4=4 >=6=6 (50% and 75%)
* 4. 18 counts: >=9=9 >=14=14 (50% and 75%)
* 5. probability splits at .5; these don't work well in prior tests
```

f2alt01d.do - 25May2002

bic' for all four outcomes in models that include all raw flim measures (fla*p5; fll*p5); pairs of u/l measures; groups of LCA measures

f2alt01e.do - all LCA probabilities - 25May2002

:::

f2alt01j.do - use three probability measures from LCA - 29May2002

:::

f2alt02c.do - 29May2002

use three binary variables, not just LC class numbers.
: dummies work better than the class number;
: effects of lower and severe are not significantly different.

Redo f2 analyses - error in adlsum - 3Jun2002

ARGH! adlsum is incorrect -- it included going to bed twice.
All of the f2alt analyses need to be redone using the corrected dataset.

f3alt_qflim07.do: create qflim07.dta 3Jun2002

1) Correct aldsum: adlsum95b
2) Add binary indicators of Lmaxp5: LmaxNonep5, etc.

f3alt01a (redo f2alt01a.do) - 3Jun2002

f3alt01b.do (redo f2 job) - 3Jun2002

HOW?

How can we use Stata and other programs to accomplish these steps?

Documentation: Keeping track of what you have done and thought

Example of Codebooks:

This is a simple example, with only the variable name, variable label, and value labels

```
FRS Coded Variables
• male          "Gender"
• age1923       "Rs Aged 19-23"
• age2429       "Rs Aged 24-29"
• age1929       "Rs Aged 19-29"
cr.ethrace      "Race/Ethnicity"
                "White:1"
                "Black:2"
                "Hisp:3"
                "Other/2+:4"
cr.edu          "Educational Attainment"
                "<HS:1"
                "HS/GED:2"
                "S.Coll:3"
                "Bach+:4"
```


HOW?

How can we use Stata and other programs to accomplish these steps?

Do-Files

- Two goals:
 1. **Robust** do-files: write do-files that produce the same result when run at a later time or on another computer
 2. **Legible** do-files: are documented and formatted so that it is easy to understand what is being done.

HOW?

How can we use Stata and other programs to accomplish these steps?

Legible Do-Files

- **Use lots of comments**
 - `* comments`
 - `// comments`
 - `/*` and `*/` comments
 - Comments as dividers

```
*****

* Comments *

*****
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Legible Do-Files

- **Use alignment and indentation**

```
rename dev origin  
rename major jobchoice  
rename HE parented
```

VS

```
rename dev      origin  
rename major    jobchoice  
rename HE       parented
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Legible Do-Files

- **Use short lines**

```
set linesize 82
```

- **Command abbreviations**

- If you want your code to be completely legible to others, do not use command abbreviations.

- A few exceptions:

generate	gen
----------	-----

summarize	sum
-----------	-----

tabulate	tab
----------	-----

- **Be Consistent**

HOW?

How can we use Stata and other programs to accomplish these steps?

Legible Do-Files: Do-File Template

```
capture log close
log using "C:\Users\kristaw\SkyDrive Pro\Stata Logs\REQ-SLB\50pl Cohabs\REQ-SLB_50pl Cohabs_CPS-2010-2012-2014-2016_06-26-17_kkp.smcl"

* File:      REQ_SLB_50pl Cohabs_CPS-2010-2012-2014-2016_
* Project:   LAT Paper for PAA
* Task:      Determine source of discrepancy in reports of # cohabiting
* Author:    Krista K. Payne
* Date:      April 18, 2017; June 26, 2017

version 14
clear all
set linesize 80
set more off
.
.
.
log close
exit
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Naming Variables

- Never change a variable unless you give it a new name.
 - `gen newvar = 0`
 - `generate` creates a new variable.
 - Almost ALWAYS generate = 0 ...I have seen many mistakes made when people generate a new variable = “.”
 - `clonevar newvar = oldvar`
 - `clonevar` generates *newvar* as an **exact copy** of an existing variable, *varname*, with the same storage type, values, and display format as *varname*. *varname*'s **variable label, value labels, notes, and characteristics will also be copied**

HOW?

How can we use Stata and other programs to accomplish these steps?

Naming Variables

- Generally, a mnemonic naming system works best ...it is the easiest for our brains to work with.

- **Mnemonic:**

- noun*

- 1. a device such as a pattern of letters, ideas, or associations that assists in remembering something.

- adjective*

- 1. aiding or designed to aid the memory.

example: dummy/binary variable to identify males—don't name it `gender` or `sex2`, because you won't know by simply reading the `varname` that all males are coded as 1.

```
gen male = 0
```

```
replace male = 1 if sex == 1
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Naming Variables

- Try to use shorter names
 - Stata allows for 32 characters, but most Stata commands show at least 12 characters for the name, so...

Use names that are at most 12 characters long

HOW? How can we use Stata and other programs to accomplish these steps?

Naming Variables

- Use capital letters sparingly → will give more meaning when you do use them.
- Some datasets (e.g., NSFG) have variable names in ALL CAPS. Recommend you convert them to all lowercase:

```
rename *, lower
```

- For household-level variables, I'll create a suffix with the first letter of the HH id variable:

```
HH id variable is serial →  
s_numbiokds
```

Letter	Meaning	Example
B/D	Binary/Dummy variable	highschlB
N	Negatively coded scale	menhlthN
P	Positively coded scale	phsyhlthP
V	Version # for modified vars.	marstV2
X	A temporary variable	Xtemp

HOW?

How can we use Stata and other programs to accomplish these steps?

Naming Variables

- Label Variables

Every variable should have a variable label.

* Beware of truncation in output

Example:

```
label variable prtmarst "Partner's Marital Status"
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Naming Variables

- You can add **notes** to variables

Example:

```
notes prtmarst: div and sep are coded together
```

To see a variables notes type:

```
notes prtmarst
```

```
prtmarst:
```

```
1. div and sep are coded together
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Naming Variables

- **Value labels** – assign text labels to the numeric values of a variable.

Categorical variables should have value labels unless the variables has an inherent metric

- Principles for constructing value labels
 1. Keep labels short: Variable labels should be eight or fewer characters in length
 2. Include the category number
 - Include them in the syntax you type
 - Use `numlabel varname, mask(#)` add
The `mask()` option for `numlabel` controls how the values are added. The `mask(#)` option adds only numbers (e.g., `1married`); `mask(#_)` adds numbers followed by an underscore (e.g., `1_married`); and `mask(#.)` adds the values followed by a period and a space (e.g., `1. married`)
 3. Avoid special characters `. := % @ { }`
 4. Apply vertically, NOT horizontally!

HOW? How can we use Stata and other programs to accomplish these steps?

Naming Variables

- **Value labels** - Principles for constructing value labels, cont.
 4. Apply vertically, NOT horizontally! (below is taken from a Stata Technical Note)

Although we tend to show examples defining value labels using one command, such as

```
. label define answ 1 yes 2 no
```

remember that value labels may include many associations and typing them all on one line can be ungainly or impossible. For instance, if perhaps we have an encoding of 1,000 places, we could imagine typing

```
. label define fips 10060 "Anniston, AL" 10110 "Auburn, AL" 10175 "Bessemer, AL" ... 560050 "Cheyenne, WY"
```

Even in an editor, we would be unlikely to type the line correctly. The easy way to enter long value labels is to enter the codings one at a time:

```
. label define fips 10060 "Anniston, AL"  
. label define fips 10175 "Bessemer, AL", add  
...  
. label define fips 560050 "Cheyenne, WY", add
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Naming Variables

- Value labels

Example:

```
label define age1929_2c 1 "1. 19-23", modify  
label define age1929_2c 2 "2. 24-29", modify  
label value age1929_2c age1929_2c  
label variable age1929_2c "YA Age Cats."
```

HOW?

How can we use Stata and other programs to accomplish these steps?

Creating Variables

- There are four simple principles for creating new variables
 1. If a variable is new, give it a new name
 2. Verify that new variables are constructed correctly
 3. Document new variables with notes and labels
 4. Keep the source variables used to create new variables

HOW?

How can we use Stata and other programs to accomplish these steps?

Creating Variables

1. If a variable is new, give it a new name

Example: Collapse the divorced and separated categories on variable `rmarital` into one category.

- Create a new variable named `rmarital_c` OR `rmaritalC` OR `rmaritalV2`

HOW?

How can we use Stata and other programs to accomplish these steps?

Creating Variables

2. Verify that new variables are constructed correctly

Example: Create a dummy variable identifying individuals in a cohabiting relationship

```
gen cohab = 0  
replace cohab = 1 if pecohab != 0  
* (100,798 real changes made)  
tab pecohab cohab, mi
```

* This is how I check if my variable is constructed correctly...by doing a crosstab of my source variable by my new variable.

* This is one way I add comments to my do files...I paste in when changes are made.

* You can also use the `list` command... I'll often do this to investigate observations with missing data on my new variable