

Data analysis tips

Vincent Kang Fu

Department of Sociology



15 February 2006

Introduction: Goals

Avoid mistakes

One seemingly innocuous error can completely derail your entire analysis

Work efficiently

Accurately communicate your data analysis

**Your results are useless and cannot be evaluated unless you explain what you did to get them
Readers should be able to replicate your results based on your write-up**

1. Hold your original data inviolate

You must be thoroughly familiar with every step you took to get from your original data to your final model

If you change your original data you may have forgotten about the change several months down the road when you are writing up your analyses

```
. use originaldata
```

```
[do recodes and create new variables here]
```

```
. save newdata
```

2. Write programs instead of working interactively

Preserve an audit trail from your original data to your final analyses—working interactively makes it too easy to forget what you've done to your data

It may be convenient to write programs interactively, but you should always compile your statements into a program

3. **Inspect your data to confirm that your commands do what you think they do**

Stata/SPSS/SAS cannot read your mind

If, for example, you do a complicated recode, inspect the results to make sure they are what you intended

```
. recode age 0/14=0 15/19=1 20/24=2 25/29=3
30/34=4 35/39=5 40/49=6 50/90=7, gen(age5)
(830630 differences between age and age5)
. label define age5 0 "0-14" 1 "15-19" 2
"20-24" 3 "25-29" 4 "30-34" 5 "35-39" 6
"40-49" 7 "50-90"
. label values age5 age5
```

```
. list age age5 in 1/27, clean
```

	age	age5	14.	4	0-14
1.	26	25-29	15.	2	0-14
2.	27	25-29	16.	2	0-14
3.	26	25-29	17.	13	0-14
4.	38	35-39	18.	11	0-14
5.	31	30-34	19.	10	0-14
6.	10	0-14	20.	26	25-29
7.	6	0-14	21.	4	0-14
8.	5	0-14	22.	15	15-19
9.	6	0-14	23.	17	15-19
10.	4	0-14	24.	12	0-14
11.	78	50-90	25.	4	0-14
12.	8	0-14	26.	21	20-24
13.	6	0-14	27.	0	0-14

4. Break your tasks into modules that you develop independently

More efficient because you will only be testing one piece a time—you will not need to read in the entire dataset again if you simply change a recode

You may even be able to reuse code for later projects

-----start of master.do-----

*

* This is my project that does X

*

do getdata

do myrecodes

do descriptives

do models

-----end of master.do-----

5. Check that conditions are true

Use the `-assert-` command often to check that statements which should be true are indeed true

`-assert-` stops your program if the assertion is false; it does nothing if the assertion is true

Syntax:

```
assert exp [if] [in]
```

When reading in ASCII data it's easy to make typographical errors

So use the `-assert-` command to make sure that conditions which should be true actually are true

Tape	Field	Item and Code	Outline	(Continued)
Location	Name			
9-10	2	[MARMON]	Month of Marriage	
		01	... January	
		02	... February	
		03	... March	
		04	... April	
		05	... May	
		06	... June	
		07	... July	
		08	... August	
		09	... September	
		10	... October	
		11	... November	
		12	... December	

```
. infix using cpmarr2.dct, using(cpmarr.dat)  
<OUTPUT OMITTED>
```

```
. assert marmon >= 1 & marmon <= 12
```

```
. assert marmon > 12  
assertion is false  
r(9);
```

-assert- will stop the execution of your program so that you can find the error and fix it

6. Document your programs

You may be young and still have a good memory, but you will be thankful for the documentation if you ever have to go back to a program you wrote months ago

This can easily happen when you need to respond to journal reviewers' comments

You will save yourself much anguish if you include notes in your programs about complicated decisions you make

*
* Spring 2006 Sociology 5965
* Demographic methods
*
* Produce data for Schoen harmonic
* mean marriage model example
* Use 1990 Census and NCHS marriage
* license data
*
* Vincent Kang Fu
* 13 February 2006 created
*

7. The `-label data-` command is handy

Attach a descriptive label to your data that Stata prints out whenever you `-use-` the dataset

```
. use "Lecture 07 marriage CA NCHS.dta"  
  
. label data "1989-1995 NCHS Marriage Licenses"  
  
. save, replace  
file Lecture 07 marriage CA NCHS.dta saved  
  
. use "Lecture 07 marriage CA NCHS.dta"  
(1989-1995 NCHS Marriage Licenses)
```

8. Missing values in Stata are larger than valid values

This can lead to subtle errors

```
. logit college momed daded siblings if age > 25
```

This command will include respondents who are missing on age, which you do not want