Introduction to SAS and Stata: Data Construction

Hsueh-Sheng Wu

CFDR Workshop Series

October 22, 2012

BGSU Center for Family and Demographic Research

Outline

- What are data?
- The interface of SAS and Stata
- Important differences between SAS and Stata
- SAS and Stata Operators
- The tasks of Data Management
- SAS and Stata commands for Data Construction
- Tips for using SAS and Stata
- Conclusion

What Are Data?

Raw data:

```
AMC Concord        40992232.5112930186401213.579999923706050
AMC Pacer           4749173  3113350173402582.529999971389770
AMC Spirit          379922   3122640168351213.079999923706050
Buick Century      48162034.516325019640196 2.93000006675720
Buick Electra       7827154  4204080222433502.410000085830690
.
.
.
VW Dasher          71402342.512216017236 973.740000009536741
VW Diesel           5397415  315204015535 903.779999971389771
VW Rabbit           4697254  315193015535 893.779999971389771
VW Scirocco         6850254  216199015636 973.779999971389771
Volvo 260         119951752.5143170193371632.980000019073491
```

Final data:

# of observation	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4099	22	3	2.5	11	2930	186	40	121	3.58	Domestic
2	AMC Pacer	4749	17	3	3	11	3350	173	40	258	2.53	Domestic
3	AMC Spirit	3799	22		3	12	2640	168	35	121	3.08	Domestic
4	Buick Century	4816	20	3	4.5	16	3250	196	40	196	2.93	Domestic
5	Buick Electra	7827	15	4	4	20	4080	222	43	350	2.41	Domestic
.	.											
.	.											
.	.											
70	VW Dasher	7140	23	4	2.5	12	2160	172	36	97	3.74	Foreign
71	VW Diesel	5397	41	5	3	15	2040	155	35	90	3.78	Foreign
72	VW Rabbit	4697	25	4	3	15	1930	155	35	89	3.78	Foreign
73	VW Scirocco	6850	25	4	2	16	1990	156	36	97	3.78	Foreign
74	Volvo 260	11995	17	5	2.5	14	3170	193	37	163	2.98	Foreign

What Are Data? (continued)

```
-------------------------------------------------------------------
          1         2         3         4         5
12345678901234567890123456789012345678901234567890123456789
-------------------------------------------------------------------
AMC Concord        40992232.5112930186401213.579999923706050
AMC Pacer          4749173  3113350173402582.529999971389770
AMC Spirit         379922   3122640168351213.079999923706050
Buick Century      48162034.516325019640196 2.93000006675720
Buick Electra      7827154  4204080222433502.410000085830690
.
.
.
VW Dasher          71402342.512216017236 973.740000009536741
VW Diesel          5397415  315204015535 903.779999971389771
VW Rabbit          4697254  315193015535 893.779999971389771
VW Scirocco        6850254  216199015636 973.779999971389771
Volvo 260         119951752.5143170193371632.980000019073491
-------------------------------------------------------------------
Column 1-13: Make
Column 19-22: price
Column 23-24:mpg
Column 25: rep78
Column 26-28: headroom
Column 29-30: truck
Column 31-34weight
Column 35-37: length
Column 38-39:turn
Column 40-42: displacement
```

What Are Data? (Continued)

- The final data set looks just like an Excel table.

- Each column represents a variable, except the first column that I added to indicate the number of observations in the data.

- Each row represents a observation, except the first row that I added to indicate the name of each variable,

- The purpose of data construction is to make a change or changes to this Excel table, for example,
  - You can change the value of a variable for some or all observations
  - You can change the name or attribute of a variable
  - You can add new variables, new observations, or both.

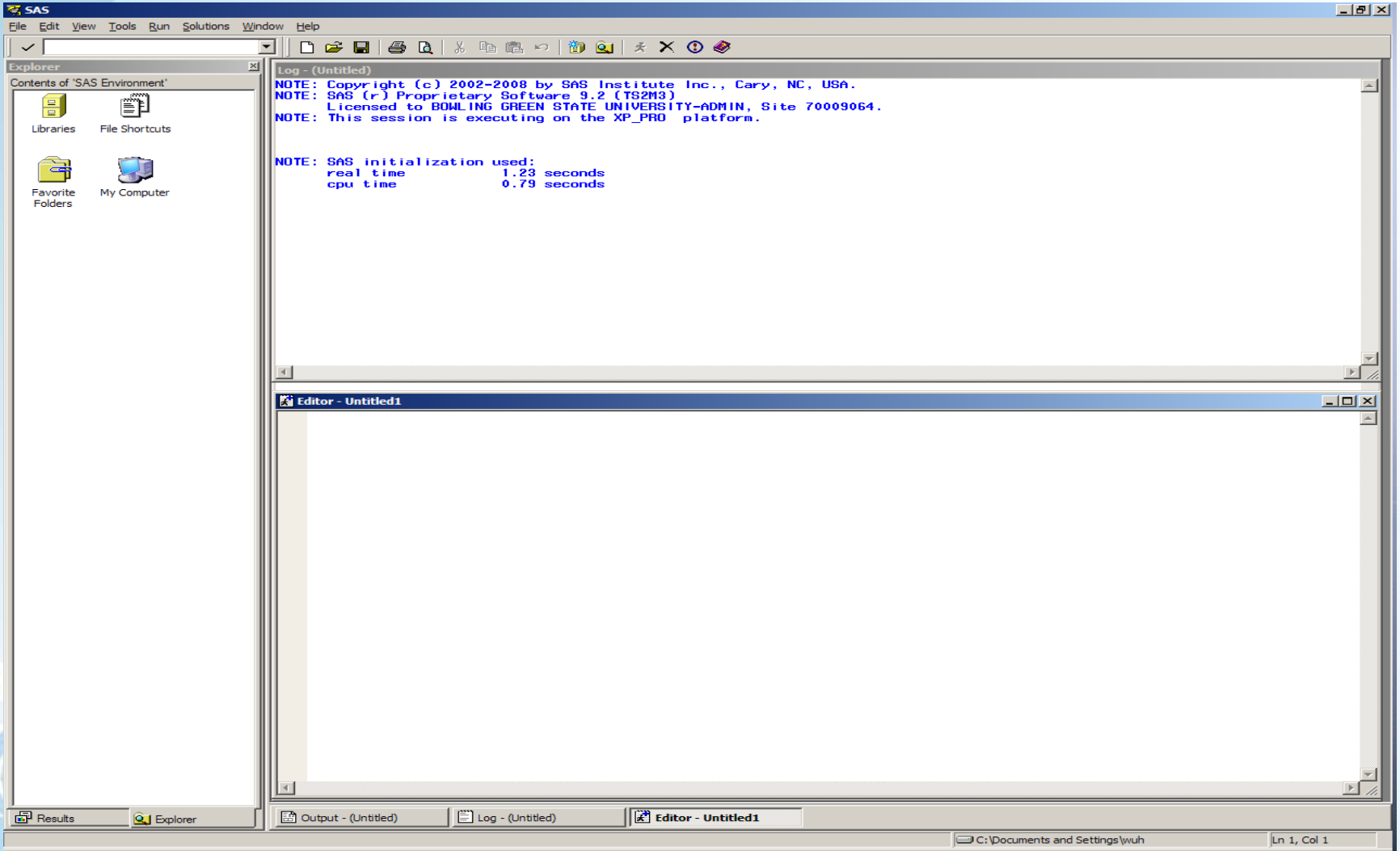
The interface of SAS and Stata

SAS user interface

- Three main windows
  - Explorer window for looking at the data
  - Editor window for writing a SAS command file
  - Log window for errors in the SAS program
- An additional window – the output window
  - The output window automatically pops up after you execute a SAS command file that produces SAS outputs
- The steps of using SAS
  - Using Editor window to write a SAS command file or
  - Execute the command file
  - Check if there are the error messages in the log window
  - Check the output in the output window
  - Remember to save your command, log, and output files

The interface of SAS and Stata (Continued)

SAS Interface



The interface of SAS and Stata (Continued)

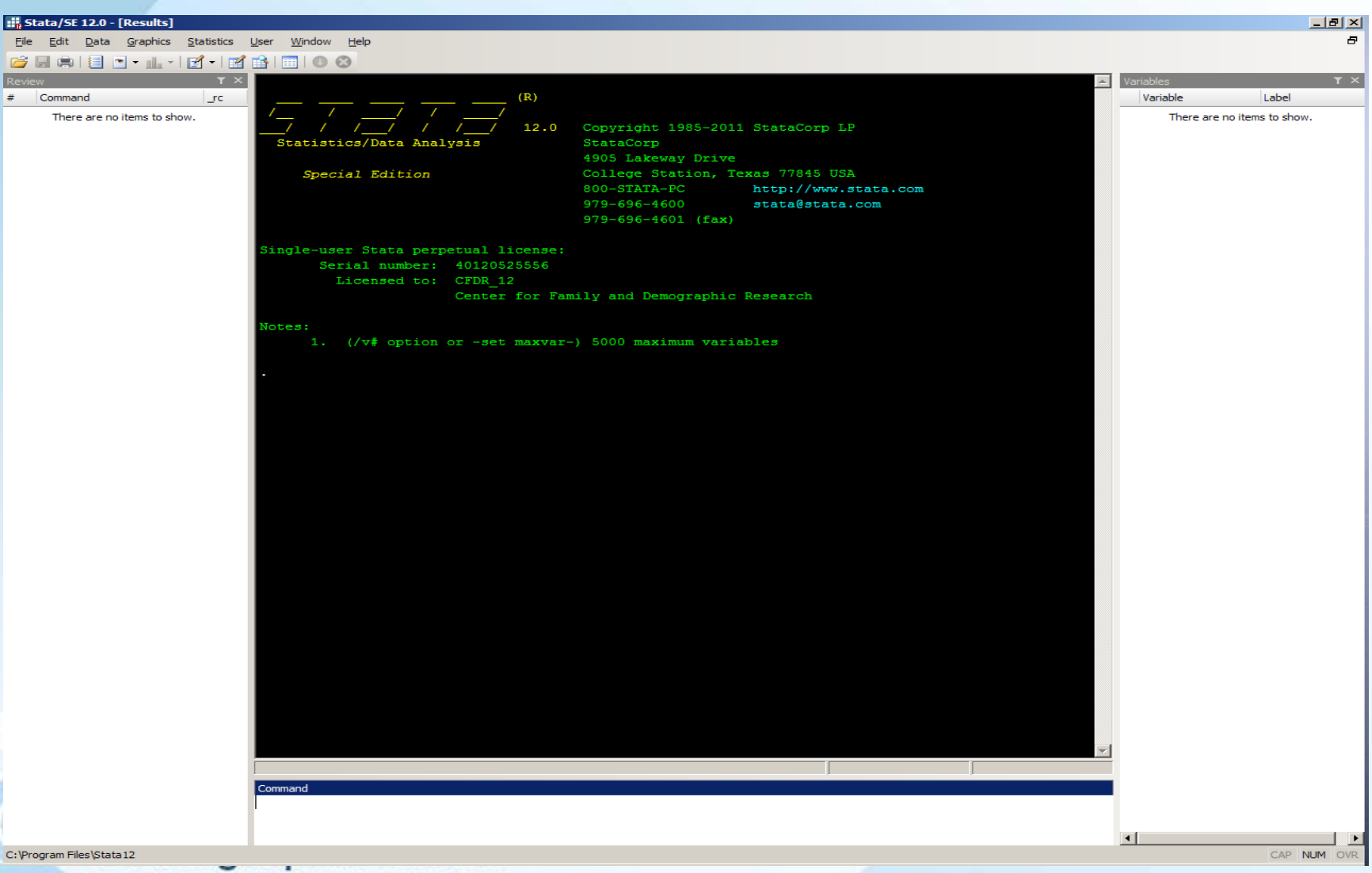
Stata user interface

- Four task windows
  - Command window: You type in the command here and press Enter to submit the command
  - Results window shows the results after commands were executed
  - Review window shows the list of executed command
  - Variables window shows the list of variables in memory

- The steps of using Stata
  - Use the new do-file editor to write the command file
  - Execute the command file
  - Check for error messages in the result window
  - Remember to save your command and log files

The interface of SAS and Stata (Continued)

Stata Interface



Important differences between SAS and Stata

- SAS reads one observation at a time, while Stata reads all observations at the same time.
- SAS commands are not case sensitive, but Stata are.
- Every SAS statement ends with a semicolon, but Stata does not.
- SAS and Stata often use different commands to achieve the same task
- Some analyses are better conducted with SAS, but some others with Stata

SAS and Stata Operators

Comparison				
	Equal	=	eq	==
	Not equal	^=	ne	~=
				!=
	Greater than	>	gt	>
	Less than	<	lt	<
	greater than or equal to	>=	ge	>=
	less than or equal to	<=	le	<=
Logical				
	both	&	and	&
	or	|	or	|
	not true	^	not	
Arithmetic				
	Addition	+		+
	subtracction	-		-
	Multiplication	*		*
	Division	/		/
	Exponentiation	**		^

SAS and Stata Operators(continued)

- The order of priorities of operators:
  - Parenthesis has a higher priority than all operators
  - All comparison Operators are equal
  - All Logical operators are equal.
  - Within Arithmetic Operators (Exponentiation > Multiplication or Division > Addition or Subtraction.
  - Among these three types of operators, Arithmetic operators > Logical operators > Comparison Operators

The Task of Data Construction

- Read in and save data
- Take a look at the data file
- Change the order of observations
- Change the order of variables
- Modify variables
- Add labels
- Create new variables
- Merge data
- Create a subset of data

Read and save data

SAS

- If you have a SAS system file (i.e., auto.sas7bdat) stored in a directory (c:\temp\in) and you want to save it to another directory (c:\temp\out)

```
LIBNAME in  "c:\temp\in";
LIBNAME out "c:\temp\out";
DATA  out.auto2;
SET  in.auto;
RUN;
```

- If you have SAS export file (i.e. auto.xpt or “auto.exp) stored in a directory (c:\temp\in) and you want to save it to another directory (c:\temp\out)

```
LIBNAME in xport "c:\temp\in\auto.xpt";
LIBNAME out "c:\temp\out";
DATA out.auto2;
SET in.auto;
RUN;
```

Read and save data (Continued)

Stata

- If you have a Stata system file (i.e., auto.dta) stored in a directory (c:\temp\in) and you want to save it to another directory (c:\temp\out)

  use “c:\temp\in\auto.dta”, clear

  save “c:\temp\out\auto.dta”, replace

- Stata data files are compatible across all platforms, so there is no portable file for Stata

Note.

- If the Data are in SPSS format, you can use Stat/Transfer to change them into SAS and Stata format.

- If you need to key in the data yourself, you can try to create them within Excel, save the file, and then use Stat/Transfer to transfer it into a SAS or Stata file.

Take a look at the data file

Find the attribute of data

SAS:

```
PROC CONTENTS DATA = in.auto position;
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
describe
```

Find summary statistics for numeric variables

SAS:

```
PROC MEANS DATA = in.auto;
VAR price mpg;
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
sum price mpg
```

Take a look at the data file

Frequencies for both numeric and string variables

SAS:

```
PROC FREQ DATA = in.auto;
TABLES make price mpg;
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
tab1 price mpg
```

Examine the values of variables for some observations

SAS

```
PROC PRINT DATA = in.auto (firstobs = 1 obs = 60);
VAR make price mpg foreign;
WHERE (mpg <=20 and foreign =0);
RUN;
```

Stata

```
use “c:\temp\in\auto.dta”, clear
list make price mpg foreign if mpg <=20 & foreign == 0 in 1/60
```

Change the order of observations

SAS:

```
PROC SORT DATA=in.auto OUT=out.auto_s;
BY mpg;
RUN;
```

Stata

```
use “c:\temp\in\auto.dta”, clear
sort mpg
save “c:\temp\out\auto.dta”, replace
```

Note: Sorting observations is important if you want to merge data files together

Change the order of variables

SAS

```
DATA out.auto2;
RETAIN      foreign make price mpg rep78
            headroom trunk weight length turn
            displacement gear_ratio;
SET in.auto;
RUN;
```

Stata

```
use “c:\temp\in\auto.dta”, clear
order       foreign make price mpg rep78 /*
*/          headroom trunk weight length turn /*
*/          displacement gear_ratio
save “c:\temp\out\auto2.dta”, replace
```

Modify variables

Rename Variables

SAS:

```
DATA out.auto2;
SET in.auto (rename=(mpg=mpg2 price=price2));
RUN;

DATA out.auto2;
SET in.auto;
RENAME mpg =mpg2 price=price2;
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
rename mpg mpg2
rename price price2
“c:\temp\out\auto2.dta”, replace
```

Modify variables

Change the value of a variable

SAS:

```
DATA out.auto2;
SET in.auto;
repair = .;
IF (rep78=1) OR (rep78=2) THEN repair = 1;
IF (rep78=3) THEN repair = 2;
IF (rep78=4) OR (rep78=5) THEN repair = 3;
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
gen repair = .
replace repair = 1 if rep78 ==1 | rep78 ==2
replace repair = 2 if rep78 = 3
replace repair = 3 if inlist(rep78, 4,5)
“c:\temp\out\auto2.dta”, replace
```

Modify variables

Change the numeric variables to string variables and vice versa

SAS:

```
DATA out.auto2;
SET in.auto;
s_mpg = put(mpg, best2.);       /* create a string variable */
n_mpg = input(s_mpg,2.0);       /* create a numeric variable */
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
tostring mpg, gen(s_mpg)       /* create a string variable */
destring s_mpg, gen(n_mpg)  /* create a numeric variable */
save “c:\temp\out\auto2.dta”, replace
```

Add labels

Add labels to the data and variables

SAS:

```
DATA out.auto2 (LABEL = "new auto data");
SET in.auto;
LABEL rep78 = "Repair Record in 1978"
      mpg = "Miles Per Gallon"
      foreign= “Foreign or Domestic car";
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
 label data  "new auto data“
 label variable rep78  "Repair Record in 1978"
 label variable mpg     "Miles Per Gallon"
 label variable foreign  “Foreign or Domestic car"
 save “c:\temp\out\auto2.dta”, replace
```

Add labels

Add and use value labels

SAS:

```
PROC FORMAT;
VALUE forgnf 0="domestic" 1="foreign" ;
VALUE $makef "AMC" ="American Motors" "Buick" ="Buick (GM)" "Cad." ="Cadillac
        (GM)" "Chev." ="Chevrolet (GM)" "Datsun" ="Datsun (Nissan)";
RUN;

PROC FREQ DATA=out.auto2;
FORMAT foreign forgnf. make $makef.;
TABLES foreign make;
RUN;
```

Stata:

```
use “c:\temp\in\auto.dta”, clear
label define forgnf 0 "domestic" 1 "foreign“
label value  foreign forgnf
tab1 foreign
save “c:\temp\out\auto2.dta”, replace
```

Create New Variables

SAS:

```
DATA out.auto2;
SET in.auto;
auto=1;
lag_mpg = lag(mpg);
If  rep78 >=3 then dummy =1;
else if rep78 <3 and rep78 ne . then dummy =0;
else dummy =.;
dummy2 = dummy*2;
interact = foreign*price;
RUN;
```

Create New Variables

Stata:

```
use “c:\temp\in\auto.dta”, clear
gen auto =1
gen lag_mpg = mpg[_n-1]
gen dummy = 1 if  rep78 >=3
replace dummy =0 if rep78 <3 & rep78 ~= .
replace dummy = . if rep78 ==.
gen dummy2 = dummy*2;
gen interact = foreign*price;
save “c:\temp\out\auto2.dta”, replace
```

Merge Data

- Before you merge data files
  - How many data files do you want to merge them together?
  - Do these data sets have variables with the same name? If they do, variables from one data file will be overwritten.
  - What ID variable or variables should be used to merge these files?
- Steps of merging data
  - Sort the first data file, based on the ID variable.
  - Sort the second data file, based on the ID variable.
  - Merge two data sets, with the use of the ID variable.

Merge Data (Continued)

- The example of one-to-one merge

Table 3 The first sample data, data1)

make	model	price	mpg	rep78	headroom
AMC	Concord	4099	22	3	2.5
AMC	Pacer	4749	17	3	3
AMC	Spirit	3799	22		3
Buick	Century	4816	20	3	4.5
Buick	Electra	7827	15	4	4
Buick	LeSabre	5788	18	3	4
Buick	Opel	4453	26		3
Buick	Regal	5189	20	3	2
Buick	Riviera	10372	16	3	3.5
Buick	Skylark	4082	19	3	3.5

Table 4. The second sample data (i.e., data2)

make	model	trunk	weight	length	turn	displacement	gear_ratio
Buick	Opel	10	2230	170	34	304	2.87
Buick	Regal	16	3280	200	42	196	2.93
Buick	Riviera	17	3880	207	43	231	2.93
Buick	Skylark	13	3400	200	42	231	3.08
Cad.	Deville	20	4330	221	44	425	2.28
Cad.	Eldorado	16	3900	204	43	350	2.19
Cad.	Seville	13	4290	204	45	350	2.24
Chev.	Chevette	9	2110	163	34	231	2.93
Chev.	Impala	20	3690	212	43	250	2.56
Chev.	Malibu	17	3180	193	31	200	2.73
Chev.	Monte Carlo	16	3220	200	41	200	2.73
Chev.	Monza	7	2750	179	40	151	2.73
Chev.	Nova	13	3430	197	43	250	2.56

Merge Data (Continued)

Expected result of one-to-one merge

Table 5. Merged data file

make	model	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio
AMC	Concord	4099	22	3	2.5						
AMC	Pacer	4749	17	3	3						
AMC	Spirit	3799	22		3						
Buick	Century	4816	20	3	4.5						
Buick	Electra	7827	15	4	4						
Buick	LeSabre	5788	18	3	4						
Buick	Opel	4453	26		3	10	2230	170	34	304	2.87
Buick	Regal	5189	20	3	2	16	3280	200	42	196	2.93
Buick	Riviera	10372	16	3	3.5	17	3880	207	43	231	2.93
Buick	Skylark	4082	19	3	3.5	13	3400	200	42	231	3.08
Cad.	Deville					20	4330	221	44	425	2.28
Cad.	Eldorado					16	3900	204	43	350	2.19
Cad.	Seville					13	4290	204	45	350	2.24
Chev.	Chevette					9	2110	163	34	231	2.93
Chev.	Impala					20	3690	212	43	250	2.56
Chev.	Malibu					17	3180	193	31	200	2.73
Chev.	Monte Carlo					16	3220	200	41	200	2.73
Chev.	Monza					7	2750	179	40	151	2.73
Chev.	Nova					13	3430	197	43	250	2.56

Merge Data (Continued)

- The example of one-to-many merge

Table 6. The Make of the Car

make	foreign
AMC	Domestic
Buick	Domestic
Cad.	Domestic
Chev.	Domestic
Dodge	Domestic
Ford	Domestic
Linc.	Domestic
Merc	Domestic
Olds.	Domestic
Plym.	Domestic
Pont.	Domestic
Audi	Foreign
BMW	Foreign
Fiat	Foreign
Honda	Foreign
Mazda	Foreign
Peugeot	Foreign
Renault	Foreign
Subaru	Foreign
Toyota	Foreign
VW	Foreign
Volvo	Foreign

Merge Data (Continued)

Table 7. The Model of the Car

make	model	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio
AMC	Concord	4099	22	3	2.5	11	2930	186	40	121	3.58
AMC	Pacer	4749	17	3	3	11	3350	173	40	258	2.53
AMC	Spirit	3799	22		3	12	2640	168	35	121	3.08
Buick	Century	4816	20	3	4.5	16	3250	196	40	196	2.93
Buick	Electra	7827	15	4	4	20	4080	222	43	350	2.41
Buick	LeSabre	5788	18	3	4	21	3670	218	43	231	2.73
.											
.											
.											
VW	Dasher	7140	23	4	2.5	12	2160	172	36	97	3.74
VW	Diesel	5397	41	5	3	15	2040	155	35	90	3.78
VW	Rabbit	4697	25	4	3	15	1930	155	35	89	3.78
VW	Scirocco	6850	25	4	2	16	1990	156	36	97	3.78
Volvo	260	11995	17	5	2.5	14	3170	193	37	163	2.98

Merge Data (Continued)

The Expected result of merging the data for the makes and models of the car

Table 8. The Expected Data of the Make and Model of the Car

make	model	price	mpg	rep78	headroom	trunk	weight	length	turn	isplacemei	gear_ratio	foreign
AMC	Concord	4099	22	3	2.5	11	2930	186	40	121	3.58	Domestic
AMC	Pacer	4749	17	3	3	11	3350	173	40	258	2.53	Domestic
AMC	Spirit	3799	22		3	12	2640	168	35	121	3.08	Domestic
Buick	Century	4816	20	3	4.5	16	3250	196	40	196	2.93	Domestic
Buick	Electra	7827	15	4	4	20	4080	222	43	350	2.41	Domestic
Buick	LeSabre	5788	18	3	4	21	3670	218	43	231	2.73	Domestic
Buick	Opel	4453	26		3	10	2230	170	34	304	2.87	Domestic
Buick	Regal	5189	20	3	2	16	3280	200	42	196	2.93	Domestic
Buick	Riviera	10372	16	3	3.5	17	3880	207	43	231	2.93	Domestic
Buick	Skylark	4082	19	3	3.5	13	3400	200	42	231	3.08	Domestic
Cad.	Deville	11385	14	3	4	20	4330	221	44	425	2.28	Domestic
Cad.	Eldorado	14500	14	2	3.5	16	3900	204	43	350	2.19	Domestic
Cad.	Seville	15906	21	3	3	13	4290	204	45	350	2.24	Domestic
.												
.												
.												
VW	Dasher	7140	23	4	2.5	12	2160	172	36	97	3.74	Foreign
VW	Diesel	5397	41	5	3	15	2040	155	35	90	3.78	Foreign
VW	Rabbit	4697	25	4	3	15	1930	155	35	89	3.78	Foreign
VW	Scirocco	6850	25	4	2	16	1990	156	36	97	3.78	Foreign
Volvo	260	11995	17	5	2.5	14	3170	193	37	163	2.98	Foreign

Merge Data (Continued)

One-to-one merge

SAS

```
PROC SORT DATA=in.data1;
BY make model;
RUN;

PROC SORT DATA=in.data2;
BY make model;
RUN;

DATA in.merged_data;
MERGE in.data1 (IN=data1)  in.data2 (IN=data2);
BY make model;
RUN;
```

Merge Data (Continued)

Stata

```
use “c:\temp\in\data1.dta”, clear
sort make model
save “c:\temp\in\data1.dta”, replace

use “c:\temp\in\data2.dta”, clear
sort make model
save “c:\temp\in\data2.dta”, replace

use “c:\temp\in\data1.dta”, clear
merge 1:1 make model using “c:\temp\in\data2.dta”
save “c:\temp\in\ new.merged_dta”, replace
```

Merge Data (Continued)

One-to-many merge

SAS

```
PROC SORT DATA=in.data3;
BY make;
RUN;

PROC SORT DATA=in.data4;
BY make;
RUN;

DATA out.merged_data;
MERGE in.data3 (IN=data3)  in.data4 (IN=data4);
BY make;
RUN;
```

Merge Data (Continued)

Stata

```
use “c:\temp\in\data3dta”, clear
sort make
save “c:\temp\in\data3.dta”, replace

use “c:\temp\in\data4.dta”, clear
sort make
save “c:\temp\in\data4.dta”, replace

use “c:\temp\in\data3.dta”, clear
merge 1:m make using “c:\temp\in\data4.dta”
save “c:\temp\out\ new.merged_dta”, replace
```

Create a Subset of Data

Keep certain variables

SAS:

```
DATA out.auto2;
SET in.auto;
KEEP make mpg;
RUN;
```

Stata

```
use “c:\temp\in\auto2.dta”, clear
keep make mpg
save “c:\temp\out\auto2.dta”, replace
```

Create a Subset of Data

Delete certain variables

SAS

```
DATA out.auto2;
SET  in.auto;
DROP make mpg;
RUN;
```

Stata

```
use “c:\temp\in\auto2.dta”, clear
drop make mpg
save “c:\temp\out\auto2.dta”, replace
```

Create a Subset of Data (Cont.)

- Keep certain respondents

```
DATA out.auto2;
SET  in.auto;
IF REP78 ^= . ;
RUN;
```

Stata:

```
use “c:\temp\in\auto2.dta”, clear
keep if rep78 ~=.
save “c:\temp\out\auto2.dta”, replace
```

Create a Subset of Data (Cont.)

Delete respondents

SAS:

```
DATA out.auto2;
SET  in.auto;
IF REP78 = .  THEN DELETE;
RUN;
```

Stata

```
use “c:\temp\in\auto2.dta”, clear
drop if rep78 ==.
save “c:\temp\out\auto2.dta”, replace
```

Tips for using SAS and Stata

- Never overwrite the original data files that CFDR or NCFMR stored on the server because the accuracy of many people's research depends on these data files.
- Always write a command file to construct the data and run the analysis. When you have command files, it is easier to keep track of what you have done and what goes wrong. You can use /* */ in both SAS and Stata to add comments.
- Also, you should save the output and log files of your analyses.
- Try to divide the data construction and analysis into different small command files. Thus, you can check the accuracy of each one of them and then combine them together.
- Try to document reasons, decision rules, and concerns in your command files. So, you know why you write certain codes.
- After creating a new variable or a subset of data, you should always check the number of observations and the frequency of the variable to make sure that your data construction is correct.

Conclusions

- SAS and Stata can achieve the same data construction tasks, although often through different commands.
- How to choose between SAS and Stata?
  - The size of data file
  - The type of data management
  - The analyses to be conducted
  - Your familiarity with the software

- Other resources for learning SAS
  - http://www.ats.ucla.edu/stat/sas/
  - http://www.cpc.unc.edu/research/tools/data_analysis/sastopics
- Other resources for learning Stata
  - http://www.ats.ucla.edu/stat/stata/
  - http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial
- CFDR programming support
  - Hsueh-Sheng Wu @ 372-3119 or wuh@bgsu.edu