Introduction to SAS and Stata: Data Construction

Hsueh-Sheng Wu
CFDR Workshop Series
February 2, 2015
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:

<table>
<thead>
<tr>
<th>make</th>
<th>price</th>
<th>mpg</th>
<th>rep78</th>
<th>headroom</th>
<th>trunk</th>
<th>weight</th>
<th>length</th>
<th>turn</th>
<th>displacement</th>
<th>gear_ratio</th>
<th>foreign</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMC Concord</td>
<td>40992232</td>
<td>5112</td>
<td>3018</td>
<td>401213</td>
<td>579999923</td>
<td>706050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC Pacer</td>
<td>4749173</td>
<td>3113</td>
<td>30173</td>
<td>342582</td>
<td>529999971</td>
<td>389770</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC Spirit</td>
<td>379922</td>
<td>3122</td>
<td>301683</td>
<td>351213</td>
<td>079999923</td>
<td>706050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick Century</td>
<td>48162034</td>
<td>5163</td>
<td>250196</td>
<td>40196</td>
<td>2.930000006</td>
<td>675720</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick Electra</td>
<td>7827154</td>
<td>4204</td>
<td>8022</td>
<td>2433502</td>
<td>410000085</td>
<td>830690</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>VW Dasher</td>
<td>71402342</td>
<td>5122</td>
<td>216017236973.740000009536741</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VW Diesel</td>
<td>5397415</td>
<td>315204015535</td>
<td>903.779999971389771</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VW Rabbit</td>
<td>4697254</td>
<td>315193015535</td>
<td>893.779999971389771</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VW Scirocco</td>
<td>6850254</td>
<td>216199015636973.779999971389771</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volvo 260</td>
<td>119951752</td>
<td>5143170193371632.980000019073491</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Final data:

<table>
<thead>
<tr>
<th># of observation</th>
<th>make</th>
<th>price</th>
<th>mpg</th>
<th>rep78</th>
<th>headroom</th>
<th>trunk</th>
<th>weight</th>
<th>length</th>
<th>turn</th>
<th>displacement</th>
<th>gear_ratio</th>
<th>foreign</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AMC Concord</td>
<td>4099</td>
<td>22</td>
<td>3</td>
<td>2.5</td>
<td>11</td>
<td>2930</td>
<td>186</td>
<td>40</td>
<td>121</td>
<td>3.58</td>
<td>Domestic</td>
</tr>
<tr>
<td>2</td>
<td>AMC Pacer</td>
<td>4749</td>
<td>17</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>3350</td>
<td>173</td>
<td>40</td>
<td>258</td>
<td>2.53</td>
<td>Domestic</td>
</tr>
<tr>
<td>3</td>
<td>AMC Spirit</td>
<td>3799</td>
<td>22</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>2640</td>
<td>168</td>
<td>35</td>
<td>121</td>
<td>3.08</td>
<td>Domestic</td>
</tr>
<tr>
<td>4</td>
<td>Buick Century</td>
<td>4816</td>
<td>20</td>
<td>3</td>
<td>4.5</td>
<td>16</td>
<td>3250</td>
<td>196</td>
<td>40</td>
<td>196</td>
<td>2.93</td>
<td>Domestic</td>
</tr>
<tr>
<td>5</td>
<td>Buick Electra</td>
<td>7827</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>20</td>
<td>4080</td>
<td>222</td>
<td>43</td>
<td>350</td>
<td>2.41</td>
<td>Domestic</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>70</td>
<td>VW Dasher</td>
<td>7140</td>
<td>23</td>
<td>4</td>
<td>2.5</td>
<td>12</td>
<td>2160</td>
<td>172</td>
<td>36</td>
<td>97</td>
<td>3.74</td>
<td>Foreign</td>
</tr>
<tr>
<td>71</td>
<td>VW Diesel</td>
<td>5397</td>
<td>41</td>
<td>5</td>
<td>3</td>
<td>15</td>
<td>2040</td>
<td>155</td>
<td>35</td>
<td>90</td>
<td>3.78</td>
<td>Foreign</td>
</tr>
<tr>
<td>72</td>
<td>VW Rabbit</td>
<td>4697</td>
<td>25</td>
<td>4</td>
<td>3</td>
<td>15</td>
<td>1930</td>
<td>155</td>
<td>35</td>
<td>89</td>
<td>3.78</td>
<td>Foreign</td>
</tr>
<tr>
<td>73</td>
<td>VW Scirocco</td>
<td>6850</td>
<td>25</td>
<td>4</td>
<td>2</td>
<td>16</td>
<td>1990</td>
<td>156</td>
<td>36</td>
<td>97</td>
<td>3.78</td>
<td>Foreign</td>
</tr>
<tr>
<td>74</td>
<td>Volvo 260</td>
<td>11995</td>
<td>17</td>
<td>5</td>
<td>2.5</td>
<td>14</td>
<td>3170</td>
<td>193</td>
<td>37</td>
<td>163</td>
<td>2.98</td>
<td>Foreign</td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------</td>
<td>-------------------</td>
<td>----------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC Concord</td>
<td>40992232.5112930186401213.579999923706050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC Pacer</td>
<td>4749173 3113350173402582.529999971389770</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC Spirit</td>
<td>379922 3122640168351213.079999923706050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick Century</td>
<td>48162034.516325019640196 2.93000006675720</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick Electra</td>
<td>7827154 4204080222433502.410000085830690</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VW Dasher</td>
<td>71402342.512216017236 973.740000009536741</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VW Diesel</td>
<td>5397415 315204015535 903.779999971389771</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VW Rabbit</td>
<td>4697254 315193015535 893.779999971389771</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VW Scirocco</td>
<td>6850254 216199015636 973.779999971389771</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volvo 260</td>
<td>119951752.5143170193371632.980000019073491</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 an 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
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

<table>
<thead>
<tr>
<th>Categories</th>
<th>Function</th>
<th>Operators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SAS</td>
</tr>
<tr>
<td><strong>Comparison</strong></td>
<td>equal</td>
<td>eq</td>
</tr>
<tr>
<td></td>
<td>not equal</td>
<td>ne</td>
</tr>
<tr>
<td></td>
<td>greater than</td>
<td>gt</td>
</tr>
<tr>
<td></td>
<td>less than</td>
<td>lt</td>
</tr>
<tr>
<td></td>
<td>greater than or equal to</td>
<td>ge</td>
</tr>
<tr>
<td></td>
<td>less than or equal to</td>
<td>le</td>
</tr>
<tr>
<td><strong>Logical</strong></td>
<td>both</td>
<td>and</td>
</tr>
<tr>
<td></td>
<td>or</td>
<td>or</td>
</tr>
<tr>
<td></td>
<td>not true</td>
<td>not</td>
</tr>
<tr>
<td><strong>Arithmetic</strong></td>
<td>addition</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>subtraction</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>multiplication</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>division</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>exponentiation</td>
<td>**</td>
</tr>
</tbody>
</table>
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)

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

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

<table>
<thead>
<tr>
<th>make</th>
<th>model</th>
<th>trunk</th>
<th>weight</th>
<th>length</th>
<th>turn</th>
<th>displacemen</th>
<th>gear_ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buick</td>
<td>Opel</td>
<td>10</td>
<td>2230</td>
<td>170</td>
<td>34</td>
<td>304</td>
<td>2.87</td>
</tr>
<tr>
<td>Buick</td>
<td>Regal</td>
<td>16</td>
<td>3280</td>
<td>200</td>
<td>42</td>
<td>196</td>
<td>2.93</td>
</tr>
<tr>
<td>Buick</td>
<td>Riviera</td>
<td>17</td>
<td>3880</td>
<td>207</td>
<td>43</td>
<td>231</td>
<td>2.93</td>
</tr>
<tr>
<td>Buick</td>
<td>Skylark</td>
<td>13</td>
<td>3400</td>
<td>200</td>
<td>42</td>
<td>231</td>
<td>3.08</td>
</tr>
<tr>
<td>Cad.</td>
<td>Deville</td>
<td>20</td>
<td>4330</td>
<td>221</td>
<td>44</td>
<td>425</td>
<td>2.28</td>
</tr>
<tr>
<td>Cad.</td>
<td>Eldorado</td>
<td>16</td>
<td>3900</td>
<td>204</td>
<td>43</td>
<td>350</td>
<td>2.19</td>
</tr>
<tr>
<td>Cad.</td>
<td>Seville</td>
<td>13</td>
<td>4290</td>
<td>204</td>
<td>45</td>
<td>350</td>
<td>2.24</td>
</tr>
<tr>
<td>Chev.</td>
<td>Chevette</td>
<td>9</td>
<td>2110</td>
<td>163</td>
<td>34</td>
<td>231</td>
<td>2.93</td>
</tr>
<tr>
<td>Chev.</td>
<td>Impala</td>
<td>20</td>
<td>3690</td>
<td>212</td>
<td>43</td>
<td>250</td>
<td>2.56</td>
</tr>
<tr>
<td>Chev.</td>
<td>Malibu</td>
<td>17</td>
<td>3180</td>
<td>193</td>
<td>31</td>
<td>200</td>
<td>2.73</td>
</tr>
<tr>
<td>Chev.</td>
<td>Monte Carlo</td>
<td>16</td>
<td>3220</td>
<td>200</td>
<td>41</td>
<td>200</td>
<td>2.73</td>
</tr>
<tr>
<td>Chev.</td>
<td>Monza</td>
<td>7</td>
<td>2750</td>
<td>179</td>
<td>40</td>
<td>151</td>
<td>2.73</td>
</tr>
<tr>
<td>Chev.</td>
<td>Nova</td>
<td>13</td>
<td>3430</td>
<td>197</td>
<td>43</td>
<td>250</td>
<td>2.56</td>
</tr>
</tbody>
</table>
Table 5. Merged data file

<table>
<thead>
<tr>
<th>make</th>
<th>model</th>
<th>price</th>
<th>mpg</th>
<th>rep78</th>
<th>headroom</th>
<th>trunk</th>
<th>weight</th>
<th>length</th>
<th>turn</th>
<th>displacement</th>
<th>gear_ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMC</td>
<td>Concord</td>
<td>4099</td>
<td>22</td>
<td>3</td>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC</td>
<td>Pacer</td>
<td>4749</td>
<td>17</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC</td>
<td>Spirit</td>
<td>3799</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick</td>
<td>Century</td>
<td>4816</td>
<td>20</td>
<td>3</td>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick</td>
<td>Electra</td>
<td>7827</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick</td>
<td>LeSabre</td>
<td>5788</td>
<td>18</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick</td>
<td>Opel</td>
<td>4453</td>
<td>26</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick</td>
<td>Regal</td>
<td>5189</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick</td>
<td>Riviera</td>
<td>10372</td>
<td>16</td>
<td>3</td>
<td>3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buick</td>
<td>Skylark</td>
<td>4082</td>
<td>19</td>
<td>3</td>
<td>3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cad.</td>
<td>Deville</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cad.</td>
<td>Eldorado</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cad.</td>
<td>Seville</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chev.</td>
<td>Chevette</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chev.</td>
<td>Impala</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chev.</td>
<td>Malibu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chev.</td>
<td>Monte Carlo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chev.</td>
<td>Monza</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chev.</td>
<td>Nova</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Merge Data (Continued)

- The example of one-to-many merge

<table>
<thead>
<tr>
<th>Table 6. The Make of the Car</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>make</td>
<td>foreign</td>
</tr>
<tr>
<td>AMC</td>
<td>Domestic</td>
</tr>
<tr>
<td>Buick</td>
<td>Domestic</td>
</tr>
<tr>
<td>Cad.</td>
<td>Domestic</td>
</tr>
<tr>
<td>Chev.</td>
<td>Domestic</td>
</tr>
<tr>
<td>Dodge</td>
<td>Domestic</td>
</tr>
<tr>
<td>Ford</td>
<td>Domestic</td>
</tr>
<tr>
<td>Linc.</td>
<td>Domestic</td>
</tr>
<tr>
<td>Merc</td>
<td>Domestic</td>
</tr>
<tr>
<td>Olds.</td>
<td>Domestic</td>
</tr>
<tr>
<td>Plym.</td>
<td>Domestic</td>
</tr>
<tr>
<td>Pont.</td>
<td>Domestic</td>
</tr>
<tr>
<td>Audi</td>
<td>Foreign</td>
</tr>
<tr>
<td>BMW</td>
<td>Foreign</td>
</tr>
<tr>
<td>Fiat</td>
<td>Foreign</td>
</tr>
<tr>
<td>Honda</td>
<td>Foreign</td>
</tr>
<tr>
<td>Mazda</td>
<td>Foreign</td>
</tr>
<tr>
<td>Peugeot</td>
<td>Foreign</td>
</tr>
<tr>
<td>Renault</td>
<td>Foreign</td>
</tr>
<tr>
<td>Subaru</td>
<td>Foreign</td>
</tr>
<tr>
<td>Toyota</td>
<td>Foreign</td>
</tr>
<tr>
<td>VW</td>
<td>Foreign</td>
</tr>
<tr>
<td>Volvo</td>
<td>Foreign</td>
</tr>
</tbody>
</table>
### Table 7. The Model of the Car

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

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


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;
Stata

use "c:\temp\in\data3.dta", 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