17.871: Solutions for Problem Set 1

- Red Text denotes stata code
- Blue text denotes emacs text
- Green text denotes stata output, e.g. tables
- /*Italics*/ denotes comments on code/output

Part I

/*Preamble*/
cd "...
log using pset1_log, replace
set more off /*useful to set this - prevents having to hit the 'more' key*/

Bob 18 95 18
Carol 21 43 27
Ted 14 67 9
Alice 12 23 31

/*Save the emacs output as “test.dat”, then...*/
infile str5 name age test1 test2 using test.dat, clear

/*Or:*/
name age test1 test2
Bob 18 95 18
Carol 21 43 27
Ted 14 67 9
Alice 12 23 31
insheet using test.dat, delimiter(" ")
compress
save "scores.dta", replace

list
+-----------------------------+
<table>
<thead>
<tr>
<th>name</th>
<th>age</th>
<th>test1</th>
<th>test2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Bob</td>
<td>18</td>
<td>95</td>
</tr>
<tr>
<td>2.</td>
<td>Carol</td>
<td>21</td>
<td>43</td>
</tr>
<tr>
<td>3.</td>
<td>Ted</td>
<td>14</td>
<td>67</td>
</tr>
<tr>
<td>4.</td>
<td>Alice</td>
<td>12</td>
<td>23</td>
</tr>
</tbody>
</table>
+-----------------------------+
Part II

use spae subset 2012.dta, clear

(1)
\texttt{tab q10 /*shows the category labels and totals for q10*/}
\texttt{tab q10, m nol /*shows the associated numerical values, including missing values.*/}

\texttt{gen longwait = .}
\texttt{replace longwait = 1 if q10==4 | q10==5}
\texttt{replace longwait = 0 if q10<4}
/*Missing values are handled by originally setting the new variable to "." and only altering it subsequently for the correct categories*/

(2)
\texttt{collapse (mean) percents=longwait (count) numbers=longwait [aw=weight], by(regstate)}

(3)
\texttt{list in 1/10}

+-------------------------------+
<table>
<thead>
<tr>
<th>regstate</th>
<th>percents</th>
<th>numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Alabama</td>
<td>.0760943</td>
<td>174</td>
</tr>
<tr>
<td>2. Alaska</td>
<td>.0222348</td>
<td>165</td>
</tr>
<tr>
<td>3. Arizona</td>
<td>.0761444</td>
<td>67</td>
</tr>
<tr>
<td>4. Arkansas</td>
<td>.1117</td>
<td>170</td>
</tr>
<tr>
<td>5. Californ</td>
<td>.0210003</td>
<td>95</td>
</tr>
<tr>
<td>6. Colorado</td>
<td>.0249651</td>
<td>67</td>
</tr>
<tr>
<td>7. Connecti</td>
<td>.0667891</td>
<td>171</td>
</tr>
<tr>
<td>8. Delaware</td>
<td>.0139119</td>
<td>190</td>
</tr>
<tr>
<td>9. District</td>
<td>.3852806</td>
<td>78</td>
</tr>
<tr>
<td>10. Florida</td>
<td>.3867461</td>
<td>131</td>
</tr>
</tbody>
</table>
+-------------------------------+

(4)
\texttt{save "wait_by_state_2012.dta", replace}
Part III

use sparesubset2014.dta, clear

(1)
gen longwait =.
replace longwait = 1 if q13==4 | q13==5
replace longwait = 0 if q13<4

collapse (mean) percents=longwait (count) numbers=longwait [aw=weight], by(inputstate)

(2)
save "wait_by_state_2014.dta",replace

(3)
rename inputstate regstate /*Variable we use for merging needs same name in both datasets*/
rename percents percents14 /*Give the variables we’re analyzing an identifying suffix*/
rename numbers numbers14
sort regstate /*Both datasets must be sorted in the same order*/
save "wait_by_state_2014.dta",replace

use "wait_by_state_2012.dta"
rename percents percents12
rename numbers numbers12
sort regstate
save "wait_by_state_2012.dta", replace

use "wait_by_state_2014.dta"
merge 1:1 regstate using "wait_by_state_2012.dta"
drop _merge

<table>
<thead>
<tr>
<th>Result</th>
<th># of obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>not matched</td>
<td>0</td>
</tr>
<tr>
<td>matched</td>
<td>51 (_merge==3)</td>
</tr>
</tbody>
</table>

/*Shows that matching was done correctly*/
(4)
list if percents14 > percents12

+------------------------------------------------------+
  regstate    perce~14   numbe~14    perce~12   numbe~12
+------------------------------------------------------+
  6.  Colorado  .0255073        31  .0249651        67
  20.   Maine    .0164273       153  .0092996       145
  32.  New Mexi  .0367988       154  .0340302       158
  38.   Oregon  .1993681         9  0  0
+------------------------------------------------------+

Part IV

use spae subset2012.dta, clear

(1)
gen longwait = .
replace longwait = 1 if q10==4|q10==5
replace longwait = 0 if q10<4
/*Leave us with percents who waited 30 mins+, for each value of q4, for each state*/
collapse (mean) percents=longwait (count) numbers=longwait [aw=weight], by(regstate q4)
/*Here's what the first ten lines look like. Notice we have multiple lines for each state*/
list in 1/10

+------------------------------------------+
  regstate    q4  percents   numbers
+------------------------------------------+
  1.  Alabama  In perso  .0760943       174
  2.   Alabama  Voted by  .  0
  3.   Alabama  I don't   .  0
  4.   Alabama   .  0
  5.  Alaska  In perso  0  136
    |------------------------------------------|
  6.  Alaska  In perso  .1432126        29
  7.  Alaska  Voted by  .  0
  8.  Alaska  I don't   .  0
  9.  Alaska   .  0
 10.  Arizona  In perso  .0881236       61
+------------------------------------------+
/*Now, re-shape to leave one observation per state, which is easier to read*/
drop if q4==.|q4>2  /*Those who didn’t vote or didn’t answer aren’t relevant here*/
reshape wide percents numbers, i(regstate) j(q4)
(note: j = 1 2)

Data    long    ->    wide

-------

Number of obs.         99    ->    51
Number of variables    4    ->    5
j variable (2 values)   q4    ->    (dropped)
xij variables:
                          percents    ->    percents1 percents2
                          numbers    ->    numbers1 numbers2

-------

/*This tells us that re-shaping split “percents” and “numbers” into two variables with the suffixes “1” and “2”. These relate to the two possible levels of q4: voting on the day, or voting early*/

(2)
list

+------------------------------+
|  regstate | percents1 | numbers1 | percents2 | numbers2 |
|-------------------------------+--------------------------------+
|  2.  Alaska   | 0        | 136      | .1432126  | 29       |
|  3.  Arizona   | .0881236 | 61       | 0         | 6        |
|  4.  Arkansas   | .0971476 | 87       | .1297159  | 83       |
|  5.  Californ   | .0213493 | 94       | 0         | 1        |
|  6.  Colorado   | .0436208 | 37       | 0         | 30       |
|  7.  Connecti   | .0546858 | 168      | .6057609  | 3        |
|  8.  Delaware   | .0140393 | 188      | 0         | 2        |
|  9.  District   | .3842785 | 129      | .3883336  | 49       |
| 10. Florida    | .3249482 | 66       | .4485258  | 65       |
| 11. Georgia    | .1564134 | 98       | .202052   | 72       |
| 12. Hawaii     | .02441   | 72       | .0976297  | 28       |
| 13. Idaho      | .0366039 | 124      | .077841   | 23       |
| 15. Indiana    | .1627559 | 132      | .0687497  | 39       |

/etc...*/
Part V

(1)
use delta_public_00_10.dta, clear

keep if academicyear==2010 /*Restrict to 2010 only*/

tab state /*Turns out there are 59 “states” featured*/
drop if state=="AS"|state=="FM"|state=="GU"|state=="MH"|state=="MP"|state=="PR"|state=="PW"|state=="VI" /*Drop observations not in the 50 states+DC. This is a good geography test*/

/*The easiest strategy to deal with missing data here is simply to drop cases that are missing one or both of revenue and enrolment.*/
keep if fte_count!=. & total03_revenue!=.

/*Now create a dataset of total revenue and total enrolment by state*/
collapse (sum) total03_revenue fte_count, by(state)

generate pcinc = total03_revenue/fte_count /*Generate Per capita revenue*/
sort pcinc /*sort the data by per capita revenue*/
list /*Shows that Massachusetts had the highest per capita revenue at $54,093 and Arizona had the lowest at $14,098*/

(2)
use delta_public_00_10.dta, clear

keep if academicyear==2008|academicyear==2007 /*Now restrict to 07 and 08*/
drop if state=="AS"|state=="FM"|state=="GU"|state=="MH"|state=="MP"|state=="PR"|state=="PW"|state=="VI"
keep state academicyear unitid total03_revenue fte_count
keep if fte_count!=. & total03_revenue!=.

/*Now create a dataset of per capita revenue by state and year*/
collapse (sum) total03_revenue fte_count, by(state academicyear)
gen pcinc = total03_revenue/fte_count
keep state academicyear pcinc
list in 1/10 /*Here’s what it looks like (first 10 rows)*/

+---------------------------------+
<table>
<thead>
<tr>
<th>academ~r</th>
<th>state</th>
<th>pcinc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2007</td>
<td>AK</td>
</tr>
<tr>
<td>2.</td>
<td>2008</td>
<td>AK</td>
</tr>
<tr>
<td>3.</td>
<td>2007</td>
<td>AL</td>
</tr>
<tr>
<td>4.</td>
<td>2008</td>
<td>AL</td>
</tr>
<tr>
<td>5.</td>
<td>2007</td>
<td>AR</td>
</tr>
<tr>
<td>6.</td>
<td>2008</td>
<td>AR</td>
</tr>
<tr>
<td>7.</td>
<td>2007</td>
<td>AZ</td>
</tr>
<tr>
<td>8.</td>
<td>2008</td>
<td>AZ</td>
</tr>
<tr>
<td>9.</td>
<td>2007</td>
<td>CA</td>
</tr>
<tr>
<td>10.</td>
<td>2008</td>
<td>CA</td>
</tr>
</tbody>
</table>
+---------------------------------+

/*It’s easier to calculate the change in revenue by first re-shaping the dataset*/
reshape wide pcinc, i(state) j(academicyear)

/*Calculate change in per capita revenue and then sort*/
gen inc_change  = pcinc2008 - pcinc2007
sort inc_change
list

/*Shows that Connecticut had the largest per capita revenue fall: a drop of $28,763*/