Introduction to Social Data Analytics

Class 13

Arushi Kaushik

arkaushi@ucsd.edu

April 30, 2019

Today: Data Wrangling in Stata

By the end of today's lecture, you should be able to:

- From pre class exercise: demonstrate appending and merging data
- Generate identifiers to differentiate between observations within a group
- Explain the difference between 1:1 and m:1 merges
- Collapse a dataset to a coarser unit of analysis
- Identify whether a dataset is long or wide and reshape it from one to the other

Open class13.do if you haven't already.

Appending vs Merging

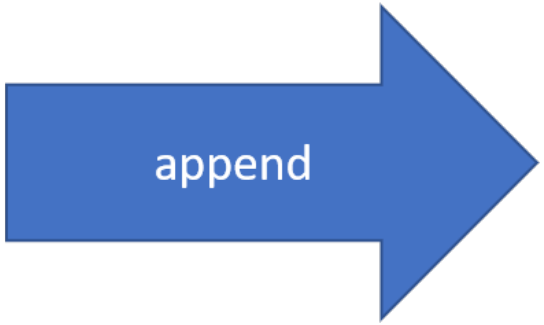
We append data to add observations, or rows.

We merge data to add variables, or columns.

Append to combine observations from tables with common variables

student	school	gpa
1	A	3.3
2	A	3.2

student	school	gpa
3	B	2.9
4	B	3.0

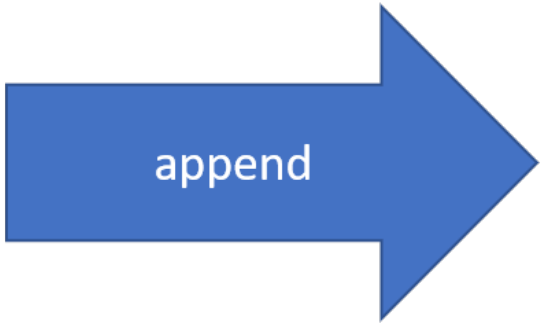


student	school	gpa
1	A	3.3
2	A	3.2
3	B	2.9
4	B	3.0

Append to combine observations from tables with common variables

student	school	gpa
1	A	3.3
2	A	3.2

student	school	gpa
3	B	2.9
4	B	3.0



student	school	gpa
1	A	3.3
2	A	3.2
3	B	2.9
4	B	3.0

From the pre class exercise: `append using person2016`

Merge 1:1 when both tables have same unit of analysis

student	school
1	A
2	A
3	B
4	B

student	gpa
1	3.3
2	3.2
3	2.9
4	3.0

merge 1:1 student

student	school	gpa
1	A	3.3
2	A	3.2
3	B	2.9
4	B	3.0

Merge 1:1 when both tables have same unit of analysis

student	school
1	A
2	A
3	B
4	B

student	gpa
1	3.3
2	3.2
3	2.9
4	3.0

merge 1:1 student

student	school	gpa
1	A	3.3
2	A	3.2
3	B	2.9
4	B	3.0

From the pre class exercise:

```
merge 1:1 year id using demographics.dta
```

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

```
sort year id
```

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

```
sort year id
```

1b. Generate an variable to indicate the first observation in each year:

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

```
sort year id
```

1b. Generate an variable to indicate the first observation in each year:

```
by year:  gen firstob = 1 if _n == 1
```

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

```
sort year id
```

1b. Generate an variable to indicate the first observation in each year:

```
by year:  gen firstob = 1 if _n == 1
```

1c. Generate means of `perwt - female` by year:

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

```
sort year id
```

1b. Generate an variable to indicate the first observation in each year:

```
by year:  gen firstob = 1 if _n == 1
```

1c. Generate means of `perwt - female` by year:

```
by year:  egen perwtmean = mean(perwt)
```

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

```
sort year id
```

1b. Generate an variable to indicate the first observation in each year:

```
by year:  gen firstob = 1 if _n == 1
```

1c. Generate means of `perwt - female` by year:

```
by year:  egen perwtmean = mean(perwt)
```

Only plot one observation per year by adding:

Next, we will generate summary statistics at the year level.

What is the unit of analysis of class13.dta?

1a. Sort your data:

```
sort year id
```

1b. Generate an variable to indicate the first observation in each year:

```
by year:  gen firstob = 1 if _n == 1
```

1c. Generate means of `perwt - female` by year:

```
by year:  egen perwtmean = mean(perwt)
```

Only plot one observation per year by adding: `if firstob == 1`

Collapse to coarsen the unit of analysis

student	school	gpa	pop
1	A	3.3	1411
2	A	3.2	1411
3	B	2.9	2692
4	B	3.0	2692

collapse (mean)
gpa pop, by(school)

school	mean_gpa	mean_pop
A	3.25	1411
B	2.95	2692

We can collapse our data to easily calculate summary statistics.

What is the unit of analysis of class13.dta?

We can collapse our data to easily calculate summary statistics.

What is the unit of analysis of class13.dta?

2a. Collapse the data from person-year to year using weights:

```
collapse (mean) perwt - female [aweight = perwt],
by(year)
```

We can collapse our data to easily calculate summary statistics.

What is the unit of analysis of class13.dta?

2a. Collapse the data from person-year to year using weights:

```
collapse (mean) perwt - female [aweight = perwt],
by(year)
```

2b. Run the code listed to rename your variables appropriately.

We can collapse our data to easily calculate summary statistics.

What is the unit of analysis of class13.dta?

2a. Collapse the data from person-year to year using weights:

```
collapse (mean) perwt - female [aweight = perwt],
by(year)
```

2b. Run the code listed to rename your variables appropriately.

2c. Save this new data frame as class13collapsed.dta:

We can collapse our data to easily calculate summary statistics.

What is the unit of analysis of class13.dta?

2a. Collapse the data from person-year to year using weights:

```
collapse (mean) perwt - female [aweight = perwt],
by(year)
```

2b. Run the code listed to rename your variables appropriately.

2c. Save this new data frame as class13collapsed.dta:

```
save class13collapsed, replace
```

We can collapse our data to easily calculate summary statistics.

What is the unit of analysis of class13.dta?

2a. Collapse the data from person-year to year using weights:

```
collapse (mean) perwt - female [aweight = perwt],
by(year)
```

2b. Run the code listed to rename your variables appropriately.

2c. Save this new data frame as class13collapsed.dta:

```
save class13collapsed, replace
```

Create your plot and notice that including weights slightly changes the mean (why?).

Merge m:1 when one table has a coarser unit of analysis

What is the unit of analysis of each table?

student	school	gpa
1	A	3.3
2	A	3.2
3	B	2.9
4	B	3.0

school	pop
A	1411
B	2692

merge m:1 school

student	school	gpa	pop
1	A	3.3	1411
2	A	3.2	1411
3	B	2.9	2692
4	B	3.0	2692

Merge m:1 when one table has a coarser unit of analysis

What is the unit of analysis of each table?

student	school	gpa
1	A	3.3
2	A	3.2
3	B	2.9
4	B	3.0

school	pop
A	1411
B	2692

merge m:1 school

student	school	gpa	pop
1	A	3.3	1411
2	A	3.2	1411
3	B	2.9	2692
4	B	3.0	2692

Your turn! Complete a m:1 merge to complete question 3.

Changing data form: "long" vs "wide"

Often we have multiple entries of a given variable for the same unit (e.g. multiple GPAs for the same student observed once per quarter).

We can present these data as long or wide and convert between the two using reshape.

student	term	gpa
1	fall	3.3
1	spring	3.2
2	fall	2.9
2	spring	3.0

reshape wide gpa, i(student) j(term)

student	fall_ gpa	spring_ gpa
1	3.3	3.2
2	2.9	3.0

Changing data form: "long" vs "wide"

Often we have multiple entries of a given variable for the same unit (e.g. multiple GPAs for the same student observed once per quarter).

We can present these data as long or wide and convert between the two using reshape.

student	term	gpa
1	fall	3.3
1	spring	3.2
2	fall	2.9
2	spring	3.0

reshape wide gpa, i(student) j(term)

student	fall_ gpa	spring_ gpa
1	3.3	3.2
2	2.9	3.0

Your turn! Reshape class13.dta to complete question 4.

Here are the commands/operators we covered today:

- `append`
- `merge 1:1; merge m:1`
- `_n`
- `collapse`
- `aweight`
- `reshape wide`
- `reshape long`