Causal Inference

Mixtape Session

Roadmap

- Hidden curriculum
  - Background
  - Empirical workflow
  - Hierarchical folder structure
  - Naming conventions
  - Version control
  - Soft skills

Missing observations

- Check the size of your dataset in Stata using `count`
- Check the number of observations per variable in Stata using `summarize`
  - → String variables will always report zero observations under `summarize` so `count if X==""` will work
- Use `tabulate` also because oftentimes missing observations are recorded with a -9 or some other illogical negative value

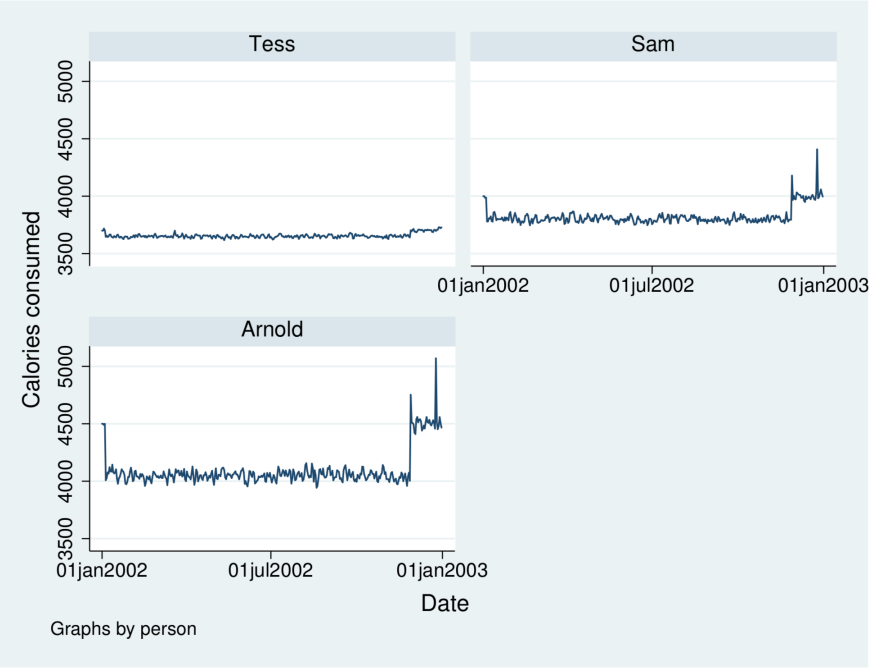
Missing years

- Panel data can be overwhelming bc looking at each state/city/firm/county borders on the impossible
- Start with `collapse` to the national level by year and simply `list` to see if anything looks strange
  - → What's "strange" look like?
  - → Well wouldn't it be strange if national unemployment rates were zero in any year?
- You can use `xtline` to see time series for panel identifiers, with or without the subcommand of `overlay`

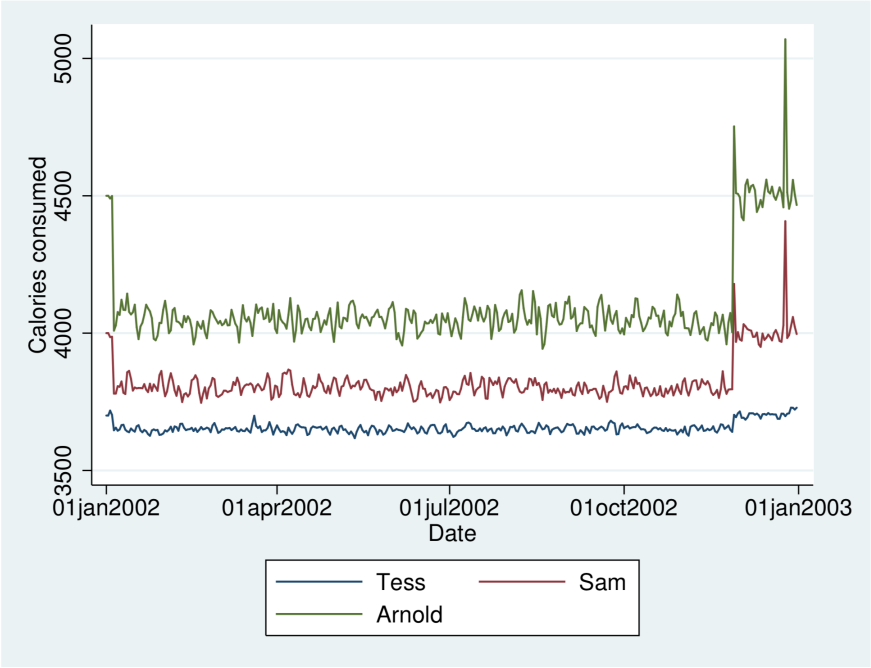
```
. collapse (sum) male_homicide female_homicide, by(year)

. list
```

	year	male_h~e	female~e
1.	1995	0	0
2.	1996	0	0
3.	1997	0	0
4.	1998	0	0
5.	1999	0	0
6.	2000	0	0
7.	2001	0	0
8.	2002	0	0
9.	2003	4474	910
10.	2004	4270	900
11.	2005	4450	895
12.	2006	4479	889
13.	2007	4480	895
14.	2008	4228	893
15.	2009	3857	866

```
. xtline calories, tlabel(#3)
```



```
. xtline calories, overlay
```



Panel observations are $N \times T$

- Say you have 51 state units (50 states plus DC) and 10 years
- $51 \times 10 = 510$ observations
- If you do not have 510 observations, then you have an unbalanced panel; if you have 510 observations you have a balanced panel
- Check the patterns using `xtdescribe` and simple counting tricks

Results

```
.
. tsset id date
       panel variable:  id (unbalanced)
        time variable:  date, 1995m1 to 2009m12
                delta:  1 month

.
.
end of do-file

. xtdescribe

      id:  4, 5, ..., 270                                    n =         69
    date:  1995m1, 1995m2, ..., 2009m12                      T =        180
           Delta(date) = 1 month
           Span(date)  = 180 periods
           (id*date uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                        24      48     180       180       180     180     180

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------------------------------------------------------------------------------------------------------------------------------------------------------------------
> ---------------
       58     84.06   84.06 |  111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
> 111111111111111
        3      4.35   88.41 |  ............................................................111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
> 111111111111111
        2      2.90   91.30 |  ............................................................................................................................................................111111111
> 111111111111111
        2      2.90   94.20 |  ........................111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
> 111111111111111
        1      1.45   95.65 |  ................................................................................................................................................111111111111111111111
> 111111111111111
        1      1.45   97.10 |  ....................................................................................................................................111111111111111111111111111111111
> 111111111111111
        1      1.45   98.55 |  ....................................................................................111111111111111111111111111111111111111111111111111111111111111111111111111111111
> 111111111111111
        1      1.45  100.00 |  111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111.............................................
> ...............
 ---------------------------+---------------------------------------------------------------------------------------------------------------------------------------------------------------------
> ---------------
       69    100.00         |  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
> XXXXXXXXXXXXXXX

.
```

log on (text)

```
. gen one = 1

. bysort county_group: egen count=sum(one)

. ta count
```

count	Freq.	Percent	Cum.
24	48	0.42	0.42
36	36	0.31	0.73
48	48	0.42	1.15
96	96	0.84	1.99
120	480	4.19	6.18
156	312	2.72	8.90
180	10,440	91.10	100.00
Total	11,460	100.00	