Stata Tutorial: Session IV

Advanced Topics

- Panel data and time series operators
- Generating professional tables
- Graphing

1) Panel Data & Time Series Operators

3 Types of Datasets:

- Cross Sectional: Many units, one observation each
  - Countries
  - Firms
  - Individuals
- Time Series: One unit, many observations each
  - One Country over many years
  - One Firm over many years (or quarters or days)
- Panel: Many units, many observations each
  - Also called “cross sectional time series”
  - “Balanced” (equal t) or “Unbalance” (different t)
  - Stata panel commands start with `xt`

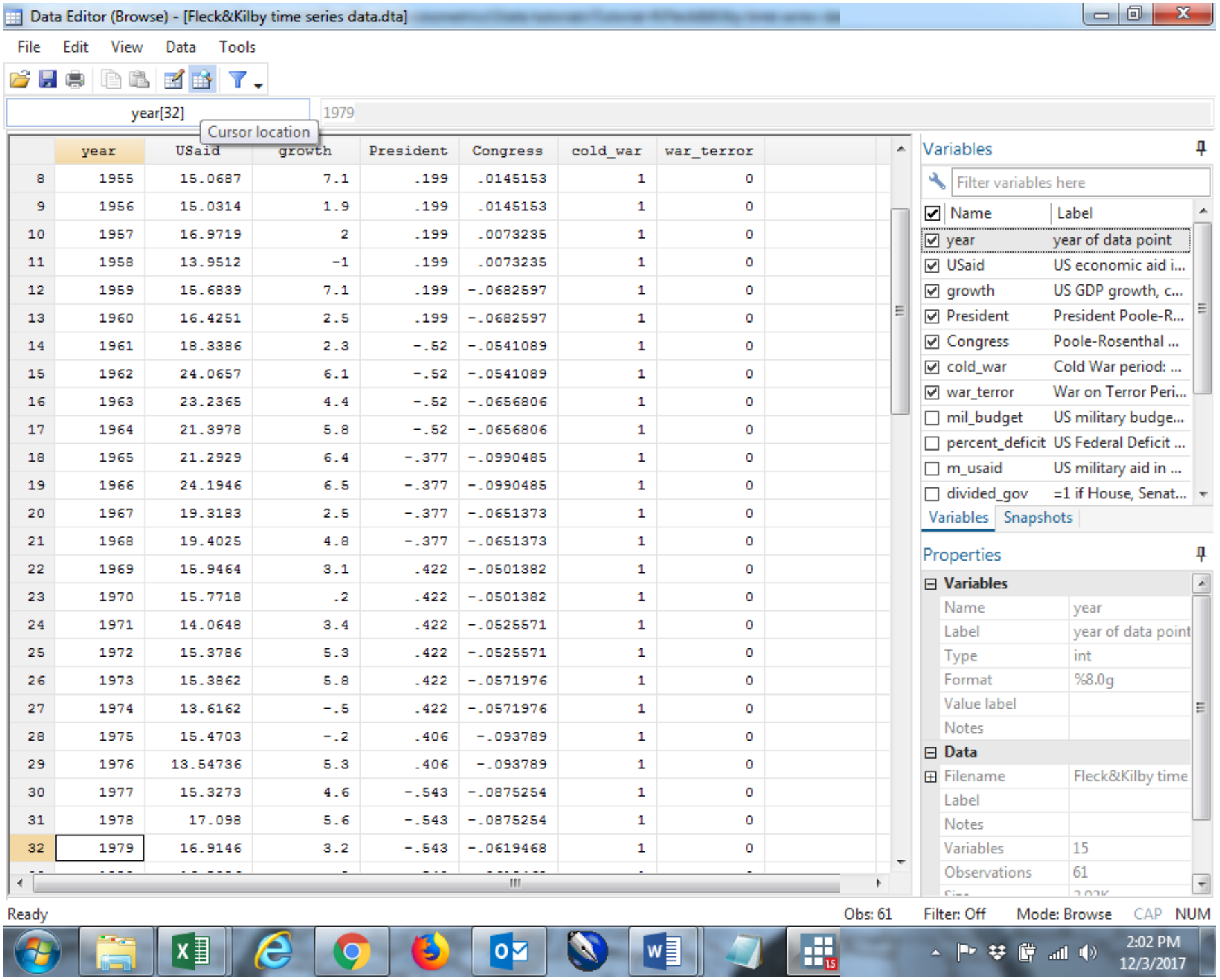
    `x` = `cross` (as in cross sectional)

    `t` = `time` (as in time series)

1) Panel Data & Time Series Operators

Why are special commands needed for Panel?

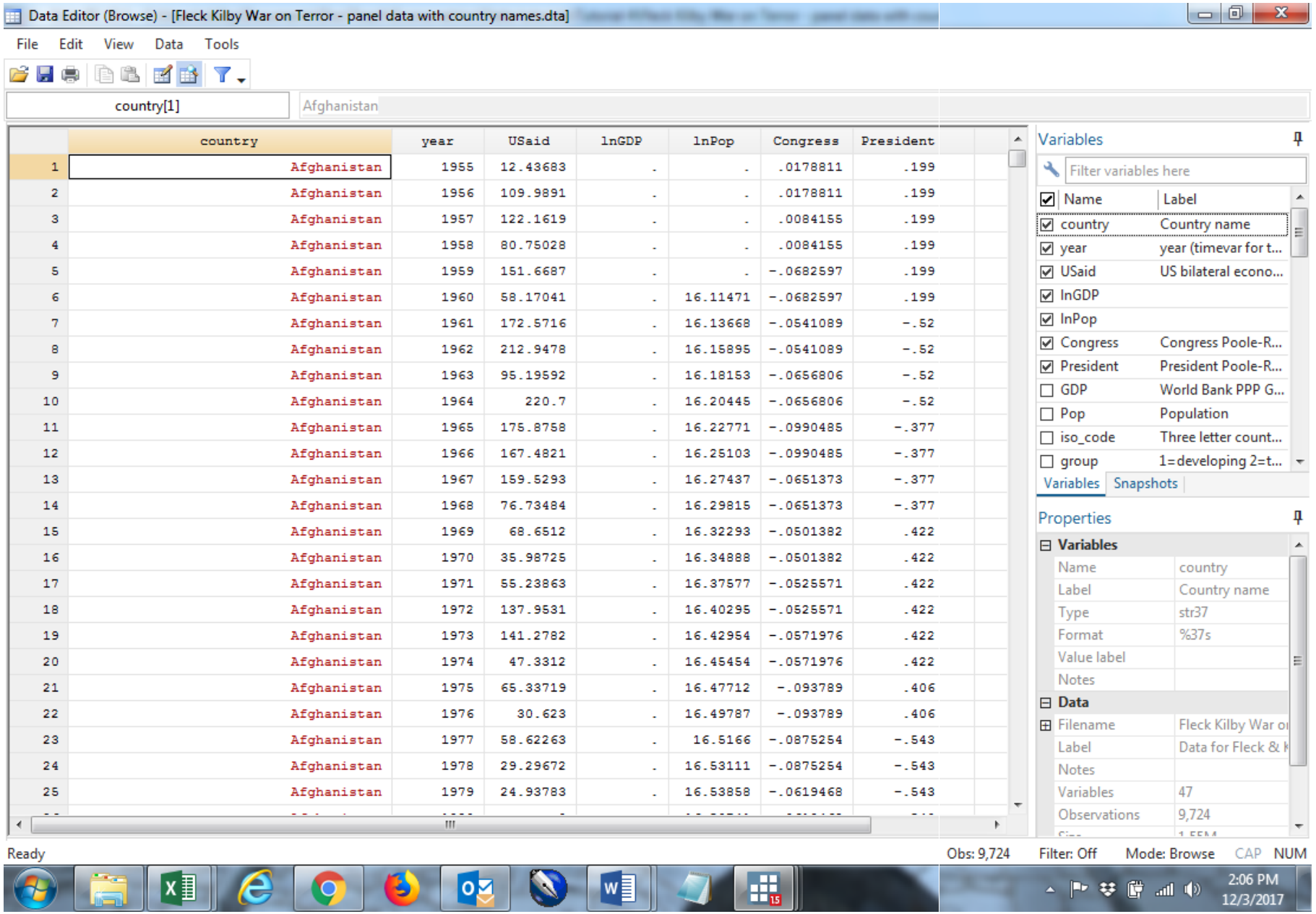
- You need to tell Stata the structure of the data
  - Time Series


Data Editor (Browse) - [Fleck&Kilby time series data.dta]
File Edit View Data Tools
year[32]
1979
Cursor location
year USaid growth President Congress cold_war war_terror
8 1955 15.0687 7.1 .199 .0145153 1 0
9 1956 15.0314 1.9 .199 .0145153 1 0
10 1957 16.9719 2 .199 .0073235 1 0
11 1958 13.9512 -1 .199 .0073235 1 0
12 1959 15.6839 7.1 .199 -.0682597 1 0
13 1960 16.4251 2.5 .199 -.0682597 1 0
14 1961 18.3386 2.3 -.52 -.0541089 1 0
15 1962 24.0657 6.1 -.52 -.0541089 1 0
16 1963 23.2365 4.4 -.52 -.0656806 1 0
17 1964 21.3978 5.8 -.52 -.0656806 1 0
18 1965 21.2929 6.4 -.377 -.0990485 1 0
19 1966 24.1946 6.5 -.377 -.0990485 1 0
20 1967 19.3183 2.5 -.377 -.0651373 1 0
21 1968 19.4025 4.8 -.377 -.0651373 1 0
22 1969 15.9464 3.1 .422 -.0501382 1 0
23 1970 15.7718 .2 .422 -.0501382 1 0
24 1971 14.0648 3.4 .422 -.0525571 1 0
25 1972 15.3786 5.3 .422 -.0525571 1 0
26 1973 15.3862 5.8 .422 -.0571976 1 0
27 1974 13.6162 -.5 .422 -.0571976 1 0
28 1975 15.4703 -.2 .406 -.093789 1 0
29 1976 13.54736 5.3 .406 -.093789 1 0
30 1977 15.3273 4.6 -.543 -.0875254 1 0
31 1978 17.098 5.6 -.543 -.0875254 1 0
32 1979 16.9146 3.2 -.543 -.0619468 1 0
Variables
Filter variables here
Name Label
year year of data point
USaid US economic aid i...
growth US GDP growth, c...
President President Poole-R...
Congress Poole-Rosenthal ...
cold_war Cold War period: ...
war_terror War on Terror Peri...
mil_budget US military budge...
percent_deficit US Federal Deficit ...
m_usaid US military aid in ...
divided_gov =1 if House, Senat...
Variables Snapshots
Properties
Variables
Name year
Label year of data point
Type int
Format %8.0g
Value label
Notes
Data
Filename Fleck&Kilby time
Label
Notes
Variables 15
Observations 61
Ready
Obs: 61
Filter: Off
Mode: Browse
CAP NUM
2:02 PM
12/3/2017

1) Panel Data & Time Series Operators

Why are special commands needed for Panel?

- You need to tell Stata the structure of the data
- Time Series
  - `tset year`
  - So Stata knows how to treat data:
    - out of order
    - gaps due to missing data
    - operations that use time
    - estimators that use time
- Panel



	country	year	USaid	lnGDP	lnPop	Congress	President
1	Afghanistan	1955	12.43683	.	.	.0178811	.199
2	Afghanistan	1956	109.9891	.	.	.0178811	.199
3	Afghanistan	1957	122.1619	.	.	.0084155	.199
4	Afghanistan	1958	80.75028	.	.	.0084155	.199
5	Afghanistan	1959	151.6687	.	.	-.0682597	.199
6	Afghanistan	1960	58.17041	.	16.11471	-.0682597	.199
7	Afghanistan	1961	172.5716	.	16.13668	-.0541089	-.52
8	Afghanistan	1962	212.9478	.	16.15895	-.0541089	-.52
9	Afghanistan	1963	95.19592	.	16.18153	-.0656806	-.52
10	Afghanistan	1964	220.7	.	16.20445	-.0656806	-.52
11	Afghanistan	1965	175.8758	.	16.22771	-.0990485	-.377
12	Afghanistan	1966	167.4821	.	16.25103	-.0990485	-.377
13	Afghanistan	1967	159.5293	.	16.27437	-.0651373	-.377
14	Afghanistan	1968	76.73484	.	16.29815	-.0651373	-.377
15	Afghanistan	1969	68.6512	.	16.32293	-.0501382	.422
16	Afghanistan	1970	35.98725	.	16.34888	-.0501382	.422
17	Afghanistan	1971	55.23863	.	16.37577	-.0525571	.422
18	Afghanistan	1972	137.9531	.	16.40295	-.0525571	.422
19	Afghanistan	1973	141.2782	.	16.42954	-.0571976	.422
20	Afghanistan	1974	47.3312	.	16.45454	-.0571976	.422
21	Afghanistan	1975	65.33719	.	16.47712	-.093789	.406
22	Afghanistan	1976	30.623	.	16.49787	-.093789	.406
23	Afghanistan	1977	58.62263	.	16.5166	-.0875254	-.543
24	Afghanistan	1978	29.29672	.	16.53111	-.0875254	-.543
25	Afghanistan	1979	24.93783	.	16.53858	-.0619468	-.543

1) Panel Data & Time Series Operators

Why are special commands needed for Panel?

- You need to tell Stata the structure of the data
- Time Series:
  - `tset year`
  - So Stata knows how to treat data:
    - out of order
    - gaps due to missing data
    - operations that use time
    - estimators that use time
- Panel:
  - `encode country, gen(cid)`
  - `xtset cid year`
  - `xtdes  /*describes time periods of panels*/`

1) Panel Data & Time Series Operators

If panel is unbalanced, can fill in missing observations

```
tsfill
```

or

```
tsfill, full
```

1) Panel Data & Time Series C

If panel is unbalanced, can

```
tsfill
```

or

```
tsfill, full
```

For each panel, fills in from lowest t to highest t.

1) Panel Data & Time Series C

If panel is unbalanced, can

```
tsfill
```

or

```
tsfill, full
```

For each panel, fills in from lowest t to highest t.

Finds the lowest and highest t's across all panels. Fills them all according to those values

1) Panel Data & Time Series Operators

If panel is unbalanced, can fill in missing observations

```
tsfill
```

or

```
tsfill, full
```

Only fills in panel id variable and time id variable.

Could fill in others via interpolation or extrapolation:

```
ipolate y x, gen(y2)  /*interpolate only*/
ipolate y x, gen(y3) epolate /*extrapolate too*/
```

Stata also can do multiple imputations with almost any command

```
help mi
```

1) Panel Data & Time Series Operators

Time Series Operators:

Once Stata knows time periods (& units, if any), can use

Lag, Forward & Difference operators

Consider regression model:

$$Y_{it} = \beta_0 + \beta_1 X_{1it-1} + \cdots + \beta_k X_{kit} + \varepsilon_{it}$$

More specifically:

$$\ln \mathrm{USaid}_{it} = \beta_0 + \beta_1 \ln \mathrm{GDP}_{it-1} + \beta_2 \ln \mathrm{Pop}_{it+1} + \varepsilon_{it}$$

Regression in Stata:

```
reg lnUSaid L.lnGDP F.lnPop
```

1) Panel Data & Time Series Operators

Time Series Operators:

$\ln \mathrm{GDP}_{it-1}$	➔	`L.lnGDP`
$\ln \mathrm{GDP}_{it-2}$	➔	`L2.lnGDP`
$\ln \mathrm{GDP}_{it-1}$ $\ln \mathrm{GDP}_{it-2}$ $\ln \mathrm{GDP}_{it-3}$	➔	`L(1/3).lnGDP`
$\ln \mathrm{GDP}_{it-1}$ $\ln \mathrm{Pop}_{it-1}$	➔	`L.lnGDP L.lnPop`
or		`L.(lnGDP lnPop)`
$\ln \mathrm{GDP}_{it+1}$	➔	`F.lnGDP`
$\ln \mathrm{GDP}_{it+1}$ $\ln \mathrm{GDP}_{it+2}$ $\ln \mathrm{GDP}_{it+3}$	➔	`F(1/3).lnGDP`
$\ln \mathrm{GDP}_{it} - \ln \mathrm{GDP}_{it-1}$	➔	`D.lnGDP`
$\ln \mathrm{GDP}_{it-1} - \ln \mathrm{GDP}_{it-2}$	➔	`LD.lnGDP`

1) Panel Data & Time Series Operators

What happens if `L.lnGDP` or `F.lnGDP` or `D.lnGDP` is missing?

- If time series, treated as missing
- If Panel, still treated as missing
  - Does not accidentally draw data from other panels

Stata tracks observations with `_n` variable

```
gen lag_lnGDP=lnGDP[_n-1]
```

is the same as

```
gen lag_lnGDP=L.lnGDP
```

except that `_n-1` does accidentally draw data from other panels

1) Panel Data & Time Series Operators

What if my data set does not have a time variable but I still want it treated as panel data? Examples

- World Bank projects by borrowing country
- Firms by industry
- People by city

Just include the panel id:

```
encode country, gen(cid)
xtset cid
```

Panel commands will work but lag operators will not.

1) Panel Data & Time Series Operators

What if I don't have just one variable that can generate a panel id but instead two or more?

Example: My panel is city/state, my time is years. But some city names appear in multiple states, e.g., Middletown, so I cannot use city alone as the panel variable.

Option #1: create new string variable with both city & state:

```
gen CityState=City+State
encode CityState, gen(csid)
```

Option #2: do it directly (but then no label on #)

```
egen csid=group(City State)
```

1) Panel Data & Time Series Operators

What if Stata complains about my variables?

```
. xtset cid year
repeated time values within panel
r(451);
```

Find repeated values using `duplicates` command:

`duplicates examples cid pid`

➔ lists examples with duplicates (often missing values…)

`duplicates tag cid pid, gen(myTag)`

➔ generates variable tagging # of duplicates

`duplicates drop`

➔ drops observations with duplicates

1) Panel Data & Time Series Operators

Can I tell Stata what my time units are? (Year, quarter, month…)

- If time variable is formatted, automatic

```
format %td year
xtset cid year
```

- Or specify in `xtset` command:

```
xtset cid year, yearly
```

1) Panel Data & Time Series Operators

Estimation and panel data

$$Y_{it} = \beta_0 + \beta_1 X_{1it-1} + \cdots + \beta_k X_{kit} + \varepsilon_{it}$$

Because we have repeated observations from same groups, very likely that within group observations are similar➔correlated!

$$E(\varepsilon_{it}\varepsilon_{is}) \neq 0 \;\leftrightarrow\; corr(\varepsilon_{it}\varepsilon_{is}) \neq 0 \;\text{ for } t \neq s$$

➔ Biased SEs, *t*-stats & *p*-values; invalid statistical inference

Usual formulas for calculating SEs assume diagonal variance-covariance matrix:

1) Panel Data & Time Series Operators

Regular SEs:

No Autocorrelation: $E(\varepsilon_{it}\varepsilon_{js}) = \sigma_\varepsilon^2$ if $i=j$ and $t=s$

$= 0$ if $i \neq j$ or $t \neq s$

➔ Homoskedastic & uncorrelated error terms

Example: Country 1, year 1,2,3; Country 2, year 1,2

$$\begin{pmatrix} \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 \end{pmatrix}$$

1) Panel Data & Time Series Operators

Clustered SEs:

Autocorrelation in country: $E(\varepsilon_{it}\varepsilon_{is}) = \sigma_i^2$ if $i=j$ and $t=s$

$= \sigma_{ii}^2$ if $i=j$ and $t \neq s$

$= 0$ if $i \neq j$

➔ Homoskedastic in country & correlated within country

$$\begin{pmatrix} \sigma_1^2 & \sigma_{11}^2 & \sigma_{11}^2 & 0 & 0 \\ \sigma_{11}^2 & \sigma_1^2 & \sigma_{11}^2 & 0 & 0 \\ \sigma_{11}^2 & \sigma_{11}^2 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_2^2 & \sigma_{22}^2 \\ 0 & 0 & 0 & \sigma_{22}^2 & \sigma_2^2 \end{pmatrix}$$

1) Panel Data & Time Series Operators

Not allowing for within panel correlation is likely to result in downward bias in estimated SEs = upward bias in t-statistics

Solution:

Pooled regression = OLS but with allowing for block diagonal Variance-Covariance matrix

Without clustering:

```
reg lnUSaid lnGDP lnPop
```

With clustering:

```
reg lnUSaid lnGDP lnPop, cluster(cid)
```

1) Panel Data & Time Series Operators

Why are error terms correlated? One reason: omitted variable!

"Unobserved Cross Sectional Heterogeneity"

$$Y_{it} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \alpha_i + \varepsilon_{it}$$

1) Panel Data & Time Series Operators

Why are error terms correlated? One reason: omitted variable!

"Unobserved Cross Sectional Heterogeneity"

$$Y_{it} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \alpha_i + \varepsilon_{it}$$

1) Panel Data & Time Series Operators

Why are error terms correlated? One reason: omitted variable!

"Unobserved Cross Sectional Heterogeneity"

$$Y_{it} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \alpha_i + \varepsilon_{it}$$

➔ Include separate dummy variable (intercept) for each panel to account for this

```
xtset cid year  /*only need to do this once*/
xtreg lnUSaid lnGDP lnPop, fe
```

➔ Can also cluster with FE though not usually as important

```
xtreg lnUSaid lnGDP lnPop, fe cluster(cid)
```

1) Panel Data & Time Series Operators

Other models:

Between Estimator: Regression on means

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_{1i} + \cdots + \beta_k \bar{X}_{ki} + \bar{\varepsilon}_i$$

```
xtreg lnUSaid lnGDP lnPop, be
```

Random Effects Estimator:

$$Y_{it} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \alpha_i + \varepsilon_{it}$$

$$\text{Assuming } \alpha_i \sim N(\alpha, \sigma_\alpha^2) \text{ and } \rho(\alpha_i, X_{jit}) = 0$$

```
xtreg lnUSaid lnGDP lnPop, re
```

1) Panel Data & Time Series Operators

Comparing Estimators:

Stata can test the validity of assumptions behind each estimator:

Do I need to use FE or is Pooled Regression OK?

➔ F-test of whether fixed effects are needed?

Last line of `xtreg` output reports this

Reject H_0 = use FE

1) Panel Data & Time Series Operators

Do I need to use FE or is RE OK?

➔ Hausman test

```
xtreg lnUSaid lnGDP lnPop, fe
eststo FE
xtreg lnUSaid lnGDP lnPop, re
eststo RE
Hausman FE RE
```

Reject H_0 = use FE

Stata has many other panel estimators – start with `xt`

```
help xt
```

2) Generating Professional Tables

Descriptive Statistics Table

- Name, # obs, Mean, SD, Min, Max

```
delimit ;
quietly reg lnUSaid lnGDP lnPop President Congress WOT;
estpost su lnUSaid lnGDP lnPop President Congress WOT if e(sample);
esttab using Table2.txt, cell((count mean sd min max)) noobs
nomtitles nonumbers tab replace;
```

variable	count	mean	sd	min	max
lnUSaid	4613	2.98153	2.074276	-2.302585	8.419823
lnGDP	4613	8.002742	.869473	6.048681	10.55457
lnPop	4613	15.66215	1.69595	10.59988	20.99467
President	4613	.0906085	.4640802	-.543	.581
Congress	4613	-.0141853	.0486689	-.0990485	.0529801
WOT	4613	.1406894	.3477385	0	1

Doesn’t work? `estimates clear`

2) Generating Professional Tables

Sample Correlations Table

- Correlations for dependent & independent variables

```
#delimit ;
quietly reg lnUSaid lnGDP lnPop President Congress WOT;
estpost corr lnUSaid lnGDP lnPop President Congress WOT if e(sample), matrix listwise;
esttab using Table3.txt, unstack not noobs nonumbers nostar nonotes compress tab replace;
```

	lnUSaid	lnGDP	lnPop	President	Congress	WOT
lnUSaid	1					
lnGDP	-0.214	1				
lnPop	0.460	-0.179	1			
President	-0.0124	0.0460	0.00150	1		
Congress	-0.0928	0.198	0.0560	0.143	1	
WOT	-0.0328	0.170	0.0420	0.331	0.488	1

2) Generating Professional Tables

Estimation Results Table

```
#delimit ;

estimates clear;

reg lnUSaid lnGDP lnPop President Congress WOT,
cluster(cid);

eststo e1;

xtreg lnUSaid lnGDP lnPop President Congress WOT, fe;

eststo e2;

esttab e1 e2 using Table4.txt, tab nogaps replace
mlabel(none) eqlabels(none) star(* .1 ** .05 *** .01)
nonotes addnotes("Dependent Variable: lnUSaid" "(1)
Pooled Regression" "(2) Country FE") title("Table 4:
Estimation Results") keep(President Congress WOT);
```

2) Generating Professional Tables

Table 4: Estimation Results

	(1)	(2)
President	0.00462	0.0132
	(0.06)	(0.30)
Congress	-4.425***	-3.839***
	(-3.78)	(-6.58)
WOT	0.112	0.382***
	(0.83)	(5.51)
N	4613	4613

Dependent Variable: lnUSaid
(1) Pooled Regression
(2) Country FE