Econometrics II: Econometric Modelling

Jürgen Meinecke

Research School of Economics, Australian National University

19 October, 2018

Logistics for week 12

We will return assignment 2 in your last tutorial

You can keep your assignment and take it with you

However, if you want to qualify for a remark you need to:

- raise and explain any concerns regarding the marking with your tutor as soon as possible during the week 12 tutorial
- hand your assignment back to your tutor by the end of the week 12 tutorial

Once you leave the tutorial room with your assignment, you cannot as for a remark

Other logistics

Tutorial participation:

- will post marks on Wattle by end of week 12
- you can contest your participation mark until 2 November (by sending me an e-mail), no remarks thereafter

Exam consults:

- will be announced in last lecture

Next Week's Lecture

Will wrap up panel data estimation in first hour

We can use second our to answer any questions you may have

Please send me any comments, suggestions and/or questions for next week

Roadmap

A **panel dataset** contains observations on multiple entities (individuals, states, companies ...), where each entity is observed at two or more points in time

Examples:

- Data on 500 Australian schools in 2010 and again in 2011, for a total of 1,000 observations
- Data on 50 U.S. states, each observed in 5 consecutive years, for a total of 250 observations
- Data on 10,000 Australians, annually, starting in 2001 (this is the famous HILDA data set)

So far we have always dealt with cross-sectional data
(ignoring our time-series adeventure in EMET2007)

Cross-sectional data have one dimension:
the unit/entity/individual i

In contrast, panel data have two dimensions:

- cross sectional: the unit/entity/individual i
- time t

Panel data notation therefore is a bit more complicated

We need a double subscript for each variable

- i = entity (state)

  n = number of entities, so $i = 1, \ldots, n$

- t = time period (year)

  T = number of time periods, so $t = 1, \ldots, T$

Typically, the cross-sectional dimension n is much larger than the time dimension T

Data: suppose we have 1 regressor; then the data are:

(X_{it}, Y_{it}) for $i = 1, \ldots, n$ and $t = 1, \ldots, T$

Notation gets even more cumbersome when we allow for k regressors

$$(X_{1it}, X_{2it}, \ldots, X_{kit}, Y_{it}) \text{ for } i = 1, \ldots, n \text{ and } t = 1, \ldots, T$$

Some jargon

- Another term for panel data is *longitudinal data*
- *balanced panel*: no missing observations, that is, all variables are observed for all entities (states) and all time periods (years)

Why are panel data useful?

Panel data help overcome problems that stem from variables that

- vary across entities i but are constant over time t
- would cause ovb if they were unobserved

Key advantage of panel data:

If an omitted variable does not change over time, then any changes in Y over time cannot be caused by that omitted variable

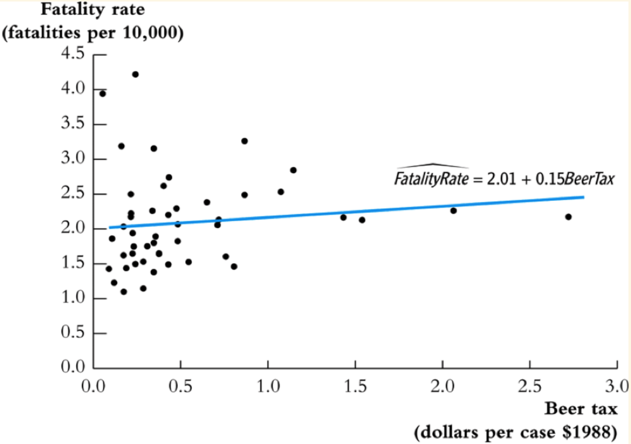
Roadmap

- Cross-sectional unit: U.S. states
  (48 U.S. states, so n = # of entities = 48)
- Time dimension: 1982 to 1988
  (7 years, so T = # of time periods = 7)
- Balanced panel, so total # observations = $7 \times 48 = 336$

Variables:

- traffic fatality rate
  (# traffic deaths in that state in that year,
  per 10,000 state residents)
- tax on a case of beer
- other: legal driving age, drunk driving laws, etc.

Just looking at 1982 data:



(a) 1982 data

Paradox: higher alcohol taxes, more traffic deaths?

Why might there be more traffic deaths in states that have higher alcohol taxes?

Other factors that determine traffic fatality rate:

- density of cars on the road
- "culture" around drinking and driving

If these are omitted, may lead to ovb

Example #1: traffic density

- High traffic density means more traffic deaths
- (Western) states with lower traffic density have lower alcohol taxes

This would lead to ovb b/c

- high taxes positively correlated with high traffic density
- high traffic density positively correlated with high fatality
- ergo: high taxes positively correlated with high fatality $\Longrightarrow$ spurious causation

Example #2: cultural attitudes towards drinking and driving

- states where people “take it easy” with drinking and driving have more traffic deaths
- these states may also have higher alcohol taxes in response to loose cultural attitudes

This would lead to ovb b/c

- high taxes positively correlated with loose cultural attitudes
- loose cultural attitudes positively correlated with high fatality
- ergo: high taxes positively correlated with high fatality $\Longrightarrow$ spurious causation

These are two hypothetical examples of how traffic fatalities could be spuriously correlated with high alcohol taxes

Let's suppose we are not able to control for traffic density and cultural attitudes

Then we may erroneously conclude that high alcohol taxes increase traffic fatalities

But these two examples (traffic density and cultural attitudes) are perfect cases for how panel data can overcome this ovb problem

Both traffic density and cultural attitudes can be thought of as being constant over time within a given entity i (here the states)

Roadmap

To get things started, suppose we only observe two time periods: 1982 and 1988

Now, consider the panel data model,

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it},$$

where Z_i is a factor that does not change over time (density), at least during the years on which we have data

- Suppose Z_i is not observed, so its omission could result in omitted variable bias
- The effect of Z_i can be eliminated using T = 2 years

The key idea:

Any change in the fatality rate from 1982 to 1988 cannot be caused by Z_i, because Z_i (by assumption) does not change between 1982 and 1988

The math: consider fatality rates in 1988 and 1982:

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}$$

$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}$$

Subtracting 1988 − 1982 (that is, calculating the change), eliminates the effect of Z_i ...

$$FatalityRate_{i1988} - FatalityRate_{i1982} =$$

$$\beta_1 (BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i1982})$$

$$FatalityRate_{i1988} - FatalityRate_{i1982} =$$
$$\beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i1982})$$

We just subtracted out the effect of Z_i

While Z_i affects the fatality rates in both years, it does not affect the *difference* in the fatality rates between the two years

The new error term, $(u_{i1988} - u_{i1982})$, is uncorrelated with either $BeerTax_{i1988}$ or $BeerTax_{i1982}$

This "difference" equation can be estimated by OLS, even though Z_i isn't observed

This differences regression doesn't have an intercept - it was eliminated by the subtraction step

Only 1982 data:

$$FatalityRate = \underset{(0.15)}{2.01} + \underset{(0.13)}{0.15} BeerTax \quad (n = 48)$$

Only 1988 data:

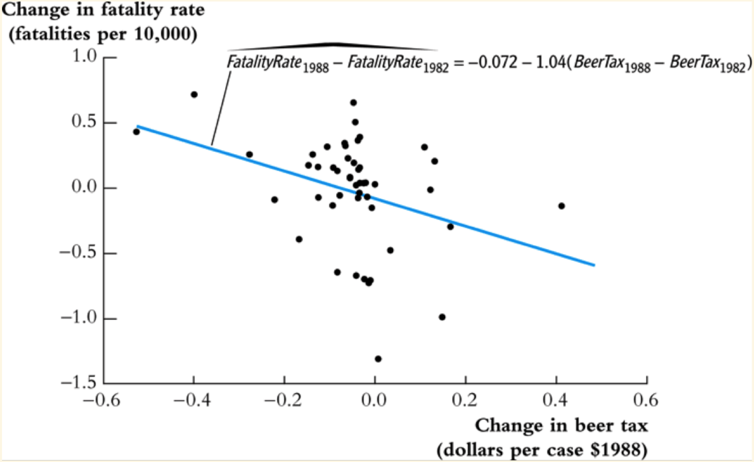
$$FatalityRate = \underset{(0.11)}{1.86} + \underset{(0.13)}{0.44} BeerTax \quad (n = 48)$$

1982 and 1988 data combined:

$$FR_{1988} - FR_{1982} = \underset{(0.065)}{-0.072} - \underset{(0.36)}{1.04} (BeerTax_{1988} - BeerTax_{1982})$$

Using the difference specification, we actually do find a *negative* effect of the beer tax on traffic fatalities (as we would expect)

Δ FatalityRate v. Δ BeerTax:



Brief summary so far:

- panel data let us remove the influence of certain unobserved variables
- any unobserved variable that does not vary over time but would affect the outcome can be taken care of
- the nice thing:
  we do not need to observe that variable at all
- we only need to exploit the time structure in a clever way and the influence of the unobserved variable disappears
- here, we only had to study the simple difference in the outcome variable between two years

Roadmap

What if you have more than 2 time periods $(T > 2)$?

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

We can rewrite this in two useful ways:

- "$n - 1$ binary regressor" regression model
- "fixed effects" regression model

We first rewrite this in "fixed effects" form

For the sake of illustration, suppose we have $n = 3$ states:
California, Texas, and Massachusetts

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

Population regression for California (that is, $i = CA$):

$$\begin{aligned} Y_{CA,t} &= \beta_0 + \beta_1 X_{CA,t} + \beta_2 Z_{CA} + u_{CA,t} \\ &= (\beta_0 + \beta_2 Z_{CA}) + \beta_1 X_{CA,t} + u_{CA,t} \end{aligned}$$

or:

$$Y_{CA,t} = \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t}$$

- $\alpha_{CA} = \beta_0 + \beta_2 Z_{CA}$ doesn't change over time
- α_{CA} is the intercept for CA, and β_1 is the slope
- The intercept is unique to CA, but the slope is the same in all the states: parallel lines

For Texas:

$$\begin{aligned} Y_{TX,t} &= \beta_0 + \beta_1 X_{TX,t} + \beta_2 Z_{TX} + u_{TX,t} \\ &= (\beta_0 + \beta_2 Z_{TX}) + \beta_1 X_{TX,t} + u_{TX,t} \\ &= \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t} \end{aligned}$$

and similarly for Massachusetts

Collecting the lines for all three states:

$$\begin{aligned} Y_{CA,t} &= \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t} \\ Y_{TX,t} &= \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t} \\ Y_{MA,t} &= \alpha_{MA} + \beta_1 X_{MA,t} + u_{MA,t} \end{aligned}$$

More compactly

$$Y_{i,t} = \alpha_i + \beta_1 X_{i,t} + u_{i,t}, \quad \text{where } i = CA, TX, MA$$

The equation

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

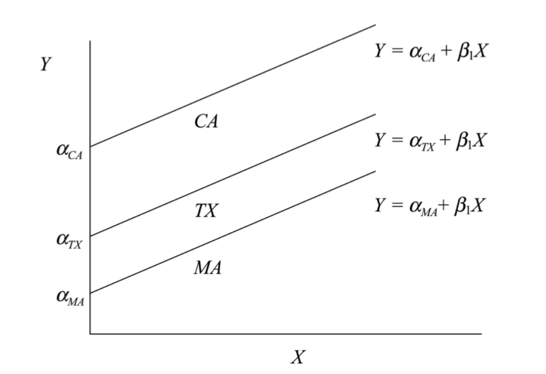
is called the *fixed effects regression model*

The terms α_i are the *state fixed effects*

All states share the same slope coefficient β_1

But each state has its own individual intercept term α_i (the fixed effect)

The regression lines for each state in a picture



The shifts in the intercept can be represented using binary regressors…

The fixed effects regression model is one of two possible representations

Alternatively, can rewrite model as *binary regressor model*

Going back to the example of three states CA, TX and MA, the fixed effects model was

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}, \quad \text{where } i = CA, TX, MA$$

It should be obvious that this can be written equivalently as

$$Y_{it} = \beta_0 + \gamma_{CA} CA_i + \gamma_{TX} TX_i + \beta_1 X_{it} + u_{it}$$

- $CA_i = 1$ if state is CA, = 0 otherwise
- $TX_i = 1$ if state is TX, = 0 otherwise

Instead of having state fixed effects α_i you include a dummy variable for each state

Summary: Two equivalent ways to write the fixed effects model

1. "Fixed effects" model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}$$

2. "n-1 binary regressor" model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_{it}$$

where

$$D2_i = \begin{cases} 1, & \text{if } i = 2 \qquad \text{(state 2)} \\ 0, & \text{otherwise} \end{cases}$$

and similary for the other dummies

Two equivalent estimation methods

1. "n-1 binary regressors" OLS regression
2. "Entity-demeaned" OLS regression

These methods produce identical estimates of the regression coefficients, and identical standard errors

Method 2 is preferred because method 1 can be computationally demanding (think of n as a large number)

“n-1 binary regressors” OLS regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_{it}$$

- First create the binary variables $D2_i, \ldots, Dn_i$
- Then simply estimate by OLS
- Inference (hypothesis tests, confidence intervals) is as usual (using heteroskedasticity-robust standard errors)
- This is impractical when n is very large
(for example if n = 1000 workers)

“Entity-demeaned” OLS regression

Entity demeaning is based on the fixed effects regression model

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

Now define entity means:

$$\bar{Y}_i := \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \bar{X}_i := \frac{1}{T}\sum_{t=1}^{T} X_{it}$$

$$\bar{u}_i := \frac{1}{T}\sum_{t=1}^{T} u_{it} \qquad \bar{\alpha}_i := \frac{1}{T}\sum_{t=1}^{T} \alpha_i = \alpha_i$$

These are the averages of each variable over time

Only for α_i: the time average coincides with the variable itself (b/c it is constant over time)

Now define "demeaned" variables:

$$\tilde{Y}_{it} := Y_{it} - \bar{Y}_i \qquad \tilde{X}_{it} := X_{it} - \bar{X}_i$$

$$\tilde{u}_{it} := u_{it} - \bar{u}_i \qquad \tilde{\alpha}_i := \alpha_i - \bar{\alpha}_i = 0$$

These are the deviations (each year) from the time average

Only for α_i: this collapses to zero
(b/c it is constant over time)

Demeaning the entire fixed effects regression model

$$\begin{aligned} Y_{it} &= \beta_1 X_{it} + \alpha_i + u_{it} \\ -\Big(\bar{Y}_i &= \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{u}_i\Big) \\ = \tilde{Y}_{it} &= \beta_1 \tilde{X}_{it} + \tilde{\alpha}_i + \tilde{u}_{it} \\ &= \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \end{aligned}$$

We have just "demeaned" the whole model

As a result, the fixed effect term α_i dropped out

This was the reason for "demeaning" in the first place

"Demeaning" is an easy data transformation that gets rid of the fixed effect (similar to first differencing in the case of $T = 2$)

Recall that when $T = 2$ we simply "first differenced" the data

The reason was that any *changes* in Y_{it} over time can only be due to *changes* in X_{it} (because the only other regressor, α_i, is fixed over time)

Now, allowing $T > 2$ we "demean" the data

The reasoning is slightly more complicated but actually quite similar:
Any deviations of Y_{it} from its time average $\bar{Y}_i$ can only be due to deviations of X_{it} from its time average $\bar{X}_i$ (because the only other regressor, α_i, is fixed over time and does not have any deviations from its time average)

Our regression model reduces to

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$$

Practical steps:

- First construct the entity-demeaned variables $\tilde{Y}_{it}$ and $\tilde{X}_{it}$
- Then simply regress $\tilde{Y}_{it}$ on $\tilde{X}_{it}$; this is straightforward OLS estimation

This is like the "changes" approach, but instead Y_{it} is deviated from the state average instead of Y_{i1}

This can be done in a single command in STATA

First let STATA know you are working with panel data by defining the entity variable (state) and time variable (year):

```
xtset state year
       panel variable:  state (strongly balanced)
        time variable:  year, 1982 to 1988
                delta:  1 unit
```

```
xtreg vfrall beertax, fe vce(cluster state)

Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48
R-sq:  within  = 0.0407                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0934                                        max =         7
                                                F(1,47)            =      5.05
corr(u_i, Xb)  = -0.6885                        Prob > F           =    0.0294

                                 (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.6558736   .2918556    -2.25   0.029    -1.243011   -.0687358
       _cons |   2.377075   .1497966    15.87   0.000     2.075723    2.678427
------------------------------------------------------------------------------
```

The panel data command *xtreg* with the option *fe* performs fixed effects regression

The *fe* option means use fixed effects regression
(Stata permits other options, but we won't discuss these)

This is equivalent to the entity demeaning described a few slides earlier

The reported intercept is arbitrary, and the estimated individual effects are not reported in the default output

The *vce(cluster state)* option tells STATA to use clustered standard errors

When you use the *vce* option, you do not need to "robustify" to heteroskedasticity (it's already being taken care of)

Stata's *xtreg* commands saves you a lot of work

- it creates the time averages $\bar{Y}_i$ and $\bar{X}_i$
- it creates the deviations of each variable from its time average, that is, $\tilde{Y}_{it}$ and $\tilde{X}_{it}$
- it runs the OLS regression of $\tilde{Y}_{it}$ on $\tilde{X}_{it}$
- it computes the correct standard errors

Why can't you use first differencing when $T > 2$?

Actually, you can

But the asymptotic properties of the estimator are not as good as the one that's based on "demeaning"

This is not at all an obvious thing but requires some nasty math to show

Bottom line: the entity demeaned estimator is more efficient than the first difference estimator

Roadmap

An omitted variable might vary over time but not across states:

- Safer cars (air bags, etc.); changes in national laws
- These produce intercepts that change over time
- Let S_t denote the combined effect of variables which changes over time but not states ("safer cars")
- The resulting population regression model is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

- Let's tentatively assume that $\beta_2 = 0$ (for sake of illustration)

This model can be recast as having an intercept that varies from one year to the next:

$$\begin{aligned} Y_{i,1982} &= \beta_0 + \beta_1 X_{i,1982} + \beta_3 S_{1982} + u_{i,1982} \\ &= (\beta_0 + \beta_3 S_{1982}) + \beta_1 X_{i,1982} + u_{i,1982} \\ &= \lambda_{1982} + \beta_1 X_{i,1982} + u_{i,1982} \end{aligned}$$

where $\lambda_{1982} = \beta_0 + \beta_3 S_{1982}$

Similarly,

$$Y_{i,1983} = \lambda_{1983} + \beta_1 X_{i,1983} + u_{i,1983}$$

where $\lambda_{1983} = \beta_0 + \beta_3 S_{1983}$

Two formulations of regression with time fixed effects

1. "Fixed effects" form
$$Y_{it} = \lambda_t + \beta_1 X_{it} + u_{it}$$

2. "T-1 binary regressor" form
$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 B2_t + \ldots + \gamma_T BT_t + u_{it}$$

   where
$$B2_t = \begin{cases} 1, & \text{if } t = 2 \quad \text{(year 2)} \\ 0, & \text{otherwise} \end{cases}$$

   and similar for other years

Two ways of estimating this

1. "Year-demeaned" OLS regression
   - Deviate Y_{it}, X_{it} from year (not state) averages
   - Estimate by OLS using "year-demeaned" data
2. "T-1 binary regressor" OLS regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 B2_t + \ldots + \gamma_T BT_t + u_{it}$$

   - Create binary variables $B2, \ldots, BT$
   - $B2 = 1$ if t = year #2, = 0 otherwise
   - Regress Y on X, B2,...,BT using OLS

Both methods result in exactly the same estimates

Because in panel data applications T is usually small, the second method is preferred here

Roadmap

What if you have both entity and time fixed effects?

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

- When T = 2, computing the first difference and including an intercept is equivalent to (gives exactly the same regression as) including entity and time fixed effects
- When T $>$ 2, there are various equivalent ways to incorporate both entity and time fixed effects:
  - entity demeaning & T - 1 time indicators (preferable and done in the following STATA example)
  - time demeaning & n - 1 entity indicators
  - T - 1 time indicators & n - 1 entity indicators
  - entity & time demeaning

```
gen y83=(year==1983)                        First generate all the time binary variables
gen y84=(year==1984)
gen y85=(year==1985)
gen y86=(year==1986)
gen y87=(year==1987)
gen y88=(year==1988)
xtreg vfrall beertax y83-y88, fe vce(cluster state)
```

```
Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48
R-sq:  within  = 0.0803                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0876                                        max =         7
corr(u_i, Xb)  = -0.6781                        Prob > F           =    0.0009
                                 (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.6399799   .3570783    -1.79   0.080    -1.358329    .0783691
         y83 |  -.0799029   .0350861    -2.28   0.027    -.1504869   -.0093188
         y84 |  -.0724206   .0438809    -1.65   0.106    -.1606975    .0158564
         y85 |  -.1239763   .0460559    -2.69   0.010    -.2166288   -.0313238
         y86 |  -.0378645   .0570604    -0.66   0.510    -.1526552    .0769262
         y87 |  -.0509021   .0636084    -0.80   0.428    -.1788656    .0770615
         y88 |  -.0518038   .0644023    -0.80   0.425    -.1813645    .0777568
       _cons |    2.42847   .2016885    12.04   0.000     2.022725    2.834215
-------------+----------------------------------------------------------------
```

Are the time effects jointly statistically significant?

```
test y83-y88

 ( 1)  y83 = 0
 ( 2)  y84 = 0
 ( 3)  y85 = 0
 ( 4)  y86 = 0
 ( 5)  y87 = 0
 ( 6)  y88 = 0

       F(  6,    47) =    4.22
            Prob > F =    0.0018
```

Yes

Practical guidelines for estimating panel data models

1. If you have individual fixed effects only

   You should use Stata's *xtreg* command

   - adding option *fe* (for fixed effect estimation)
   - adding option *vce(cluster state)* (for correct standard errors) (note: depending on your application, the cross-sectional variable "state" may have a different name)

2. If you have both individual and time fixed effects

   You should use Stata's *xtreg* command

   - adding dummy variables for $T-1$ years
   - adding option *fe* (for fixed effect estimation)
   - adding option *vce(cluster state)* (for correct standard errors) (note: depending on your application, the cross-sectional variable "state" may have a different name)

But remember:

Before you use *xtreg* you need to tell Stata that your data set is a panel data set by typing xtset state year

3. If you have time fixed effects only
   You should use Stata's *regress* command
   - adding option *vce(cluster state)* (for correct standard errors) (note: depending on your application, the cross-sectional variable "state" may have a different name)