Econometrics II: Econometric Modelling

Jürgen Meinecke

Research School of Economics, Australian National University

26 October, 2018

Logistics for week 12

We will return assignment 2 in today's 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

Uncollected assignments

I keep ALL uncollected assignments, swing by to collect yours!

If you want to be able to get a remark on an uncollected assignment 2, please collect it from me before 2 November

Any assignment 2 collected later than that cannot get a remark!

You cannot get a remark on an uncollected assignment 1 anymore

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 consultation

Final exam consultation is available on the following days:

- Friday 2 November, 11am - 12pm, Arndt 1022
- Friday 9 November, 11am - 12pm, Arndt 1022
- Monday 12 November, 10am - 12pm, CBE tutorial room 7

No other consultations are offered

We do not help with practice exams!

I will NOT use e-mail for consultation!

Final Exam

Reminder: absolutely everything is examinable!

This includes the material from today's lecture

Roadmap

Using panel data, we can extend our standard linear model to look like this:

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

(Note: including more explanatory variables is easy)

There are three error terms here: α_i, λ_t and u_{it}

As usual, error terms only pose problems to the extent that they are correlated with X_{it}

If α_i and λ_t are correlated with X_{it} then panel data will solve this problem

If u_{it} is correlated with X_{it} then panel data will not help

Under a panel data version of the least squares assumptions, the OLS fixed effects estimator of β_1 is normally distributed

However, a new standard error formula needs to be introduced: the "clustered" standard error formula

This new formula is needed because observations for the same entity are not independent (it's the same entity!)

Having said that, observations across different entities are still assumed independent

Consider the generic panel data model with entity fixed effects (for simplicity we are ignoring λ_t)

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

The four OLS assumptions adapted to the panel data model

1. $E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$
2. $(X_{i1}, \ldots, X_{iT}, Y_{i1}, \ldots, Y_{iT})$, $i = 1, \ldots, n$ are i.i.d. draws from their joint distribution
3. (X_{it}, Y_{it}) have finite fourth moments
4. heteroskedasticity

Assumption 1: $E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$

- u_{it} has mean zero, given the entity fixed effect and the entire history of the X's for that entity
- This means there are no omitted lagged effects (any lagged effects of X must enter explicitly)
- Also, there is not feedback from u to future X:
  - Whether a state has a particularly high fatality rate this year doesn't subsequently affect whether it increases the beer tax
  - Sometimes this "no feedback" assumption is plausible, sometimes it isn't

Assumption 2: $(X_{i1}, \ldots, X_{iT}, Y_{i1}, \ldots, Y_{iT})$ are i.i.d. draws from their joint distribution

- This is satisfied if entities are randomly sampled from their population by simple random sampling
- This does **not** require observations to be i.i.d. over time for the same entity
  (that would be unrealistic)
- Example: Whether a state has a high beer tax this year is a good predictor of (correlated with) whether it will have a high beer tax next year.
- Put differently, the error term for an entity in one year is plausibly correlated with its value in the year, that is, $corr(u_{it}, u_{it+1})$ is often plausibly nonzero

Under the LS assumptions for panel data

- The OLS fixed effect estimator $\hat{\beta}_1$ is unbiased, consistent, and asymptotically normally distributed
- However, the usual OLS standard errors (both homoskedasticity-only and heteroskedasticity-robust) will in general be wrong because they assume that u_{it} is serially uncorrelated
- In practice, the OLS standard errors often understate the true sampling uncertainty: if u_{it} is correlated over time, you don't have as much information (as much random variation) as you would if u_{it} were uncorrelated
- This problem is solved by using "clustered" standard errors

Clustered Standard Errors

- Clustered standard errors estimate the variance of $\hat{\beta}_1$ when the variables are i.i.d. across entities but are potentially autocorrelated within an entity
- Clustered SEs are easiest to understand if we first consider the simpler problem of estimating the mean of Y using panel data…

Let me present the following "toy model" to illustrate the main idea

$$Y_{it} = \beta_0 + u_{it},$$

A reasonable estimator for β_0 would be the sample average across entities and time:

$$\bar{Y} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it}$$

It is useful to write $\bar{Y}$ as the average across entities of the mean value for each entity:

$$\bar{Y} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{T} \sum_{t=1}^{T} Y_{it} \right) = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i$$

where $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}$ is the sample mean for entity i

Because observations are i.i.d. across entities,
$(\bar{Y}_1, \ldots, \bar{Y}_n)$ are i.i.d.

Thus, if n is large, the CLT applies and:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_i \overset{\text{approx.}}{\sim} \text{N}\left(\beta_0, \frac{\sigma^2_{\bar{Y}_i}}{n}\right), \text{ where } \sigma^2_{\bar{Y}_i} = \text{Var}\left(\bar{Y}_i\right)$$

- The SE of $\bar{Y}$ is the square root of an estimator of $\frac{\sigma^2_{\bar{Y}_i}}{n}$
- Natural estimator of $\sigma^2_{\bar{Y}_i}$ given by sample variance of $\bar{Y}_i$:
  $s^2_{\bar{Y}_i} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{Y}_i - \bar{Y})^2$

This delivers the clustered standard error formula for $\bar{Y}$
computed using panel data: $SE\left(\bar{Y}\right)_{clustered} = s_{\bar{Y}_i}/\sqrt{n}$

What's special about clustered SE?

- Not much, really - the previous derivation is the same as in EMET2007 when we derived the SE for the sample average (except that here the "data" are i.i.d. entity averages $(\bar{Y}_1, \ldots, \bar{Y}_n)$ instead of a single i.i.d. observation for each entity.
- But there is one more subtle difference:
  in the cluster SE derivation we never assumed that observations are i.i.d. within an entity
- Thus we have implicitly taken care of the possibility of serial correlation within an entity
- Where exactly did this happen?

$$SE\left(\bar{Y}\right)_{clustered} = s_{\bar{Y}_i}/\sqrt{n},$$

with
$$s^2_{\bar{Y}_i} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{Y}_i - \bar{Y})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}Y_{it} - \bar{Y}\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \bar{Y})\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \bar{Y})\right)\left(\frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \bar{Y})\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(Y_{it} - \bar{Y})(Y_{is} - \bar{Y})$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_{it} - \bar{Y})(Y_{is} - \bar{Y})\right]$$

Now compare $SE\left(\bar{Y}\right)_{clustered}$ to $SE\left(\bar{Y}\right)$

Friendly reminder from EMET2007:

$$SE\left(\bar{Y}\right) := s_Y / \sqrt{n}$$

$$\text{where} \quad s_Y^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Juxtaposing the clustered version:

$$s_{\bar{Y}_i}^2 = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_{it} - \bar{Y})(Y_{is} - \bar{Y})\right]$$

If $T = 1$, they would be the same

Generally, s_Y^2 does not include any autocovariance terms

The clustered version, $s_{\bar{Y}_i}$ on the other hand does

To see this, look at the final term in brackets:

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_{it}-\bar{Y})(Y_{is}-\bar{Y})$$

Whenever $s \neq t$, that term represents the sample autocovariance of order $|s-t|$

For example, when $s=2$ and $t=5$ then that term captures the third order autocovariance (because $5-2=3$)

Whenever $s=t$, that term represents the variance

Thus the clustered SE formula implicitly is estimating all the autocovariances, then using them to estimate $\sigma^2_{\bar{Y}_i}$

In contrast, the "usual" SE formula zeros out these autocovariances by omitting all the cross terms - which is only valid if those autocovariances are all zero

Now, extending this to panel data, the notation gets a lot more messy

But the general idea remains the same

The End

It's been a pleasure teaching you!

Really!