

## Exercise 8. Panel Data (Answers)

1. Given 
$$y_{it} = m_t + a_i + u_{it} \quad \begin{array}{l} i = 1, \dots, N \\ t = 1, \dots, T \end{array}$$

It follows that  $\text{Var}(m_t - m_s) = \text{Var}(m_t) + \text{Var}(m_s) - 2\text{Cov}(m_t, m_s)$

With panel data it is likely that  $\text{Cov}(m_t, m_s) > 0$

(the values of any variable are correlated over time for the same individuals)

With pooled data then the cross-section units are independent over time and so  $\text{Cov}(m_t, m_s) = 0$

Hence  $\text{Var}(m_t - m_s)^{\text{panel}} < \text{Var}(m_t - m_s)^{\text{pooled}}$

And so panel data estimation is likely to give more efficient estimates of **changes** (assuming no measurement error)

But pooled data is likely to give more efficient estimates of **sums or averages** since

$$\text{Var}(m_t + m_s) = \text{Var}(m_t) + \text{Var}(m_s) + 2\text{Cov}(m_t, m_s)$$

Note that the error terms from pooled estimation are positively correlated

$$\text{Cov}(e_1, e_2) = \text{Cov}(a_i + u_{i1}, a_i + u_{i2}) = \text{Var}(a_i) > 0$$

so that estimated OLS standard errors in a pooled regression are inefficient

2. 
$$y_{it} = b_0 + b_1 x_{1it} + b_2 x_{2it} + b_3 a_i + u_{it} \quad (1) \quad \begin{array}{l} i = 1, \dots, N \\ t = 1, \dots, T \end{array}$$
  
and  $u_{it} \sim \text{iid}(0, \sigma_u^2)$

a) Show that 1<sup>st</sup> differencing will introduce negative autocorrelation into the (differenced) error term with correlation coefficient  $\rho = -0.5$

Assuming (as in the question) that the original residuals are homoskedastic and uncorrelated over time, then 1<sup>st</sup> differencing (1) gives

$$\Delta y_{it} = b_1 \Delta x_{1it} + b_2 \Delta x_{2it} + \Delta u_{it} \quad \begin{array}{l} i = 1, \dots, N \\ t = 2, \dots, T \end{array}$$

which eliminates the fixed effect and gives consistent estimates as long as

$$p \lim \left( \frac{\Delta X_{kt}, \Delta u_t}{N(T-1)} \right) = p \lim \left( \frac{(X_{kt} - X_{k,t-1})(u_t - u_{t-1})}{N(T-1)} \right) = 0 \quad \text{for all } t \text{ and all } k$$

This is satisfied by assumption (the observed  $X$  values are uncorrelated with the residual) as **either**  $T$  or (more likely in a typical panel)  $N \rightarrow \infty$

[Note a similar argument applies to the consistency of the within-group estimator which relies on

$$p \lim \left( \frac{\sum_{i=1}^N \sum_{t=1}^T (X_{it} - \bar{X}_i)(u_{it} - \bar{u}_i)}{NT} \right) = 0$$

However in the 1<sup>st</sup> difference model

$$\text{Var}(\Delta u_t) = \text{Var}(u_t - u_{t-1}) = \text{Var}(u_t) + \text{Var}(u_{t-1}) - 2\text{Cov}(u_t, u_{t-1}) \quad \text{for all } i$$

Assuming that these residuals are not correlated over time then  $\text{Cov}(u_t, u_{t-1})=0$  and

$$\text{Var}(\Delta u_t) = \sigma_u^2 + \sigma_u^2 = 2\sigma_u^2$$

(constant, homoskedastic variance)

$$\text{Similarly } \text{Cov}(\Delta u_t, \Delta u_{t-1}) = E(\Delta u_t, \Delta u_{t-1})$$

(since  $E(\Delta u_t) = E(\Delta u_{t-1}) = 0$ )

$$\begin{aligned} &= E[(u_t - u_{t-1})(u_{t-1} - u_{t-2})] \\ &= E(u_t, u_{t-1}) - E[u_{t-1}^2] + E[u_{t-1}, u_{t-2}] - E[u_t, u_{t-2}] \\ &= -\sigma_u^2 \end{aligned}$$

(since no autocorrelation the 2<sup>nd</sup> term is the only non-zero term. So the form of autocorrelation is AR(1). This can be useful to know when trying to fix up the standard errors)

It follows that

$$\text{corr}(\Delta u_{it}, \Delta u_{it-1}) = \frac{\text{Cov}(\Delta u_{it}, \Delta u_{it-1})}{(\text{Var}(\Delta u_{it})\text{Var}(\Delta u_{it-1}))^{1/2}} = \frac{-\sigma_u^2}{\sqrt{2\sigma_u^2 * 2\sigma_u^2}} = \frac{-\sigma_u^2}{2\sigma_u^2} = -\frac{1}{2}$$

So 1<sup>st</sup> differencing introduces autocorrelation into the residuals which leads to inefficient standard errors

Find the degree of autocorrelation induced by within-groups estimation of this model

Within-groups estimation implies estimate

$$y_i - \bar{y}_i = (X_i - \bar{X}_i)\beta + (u_i - \bar{u}_i) \quad \text{for } i = 1, \dots, N \text{ and } t = 1, \dots, T$$

Assuming no autocorrelation or heteroskedasticity in the  $u_{it}$  error term then  $\tilde{u} = (u_i - \bar{u}_i) \sim N(0, \sigma_u^2)$

Consider the covariance between any 2 within-group residuals

$$E[\tilde{u}_{it} \tilde{u}_{is}] = E\left[ (u_{it} - \bar{u}_i)(u_{is} - \bar{u}_i) \right] = E(u_{it}u_{is}) - E(u_{it}\bar{u}_i) - E(u_{is}\bar{u}_i) + E(\bar{u}_i^2)$$

Since

$$E\left[ u_{it} \frac{\bar{u}_i}{T} \right] = E\left[ u_{it} \frac{\sum_{t=1}^T u_{it}}{T} \right] = E\left( \frac{u_{it}^2}{T} \right) = \frac{\sigma_u^2}{T}$$

(assuming original levels residuals are not correlated ie  $E(u_{it}u_{is}) = 0$  all  $s$  not equal  $t$ )

Similarly

$$E\left[\bar{u}_i^2\right] = E\left(\frac{\sum_{t=1}^T u_{it}^2}{T^2}\right) = \frac{T\sigma_u^2}{T^2} = \frac{\sigma_u^2}{T}$$

So

$$E(\tilde{u}_{it} \tilde{u}_{is}) = 0 - \frac{\sigma_u^2}{T} - \frac{\sigma_u^2}{T} + \frac{\sigma_u^2}{T} = -\frac{\sigma_u^2}{T}$$

which is also the equation for the covariance since

$$E(\tilde{u}_{it}) = E(\tilde{u}_{is}) = E(u_{it}) - E(\bar{u}_i) = E(u_{it}) - E\left(\frac{\sum_{t=1}^T u_{it}}{T}\right) = 0$$

$$\text{Hence } \text{corr}(\tilde{u}_{it} \tilde{u}_{is}) = \frac{\text{Cov}(\tilde{u}_{it} \tilde{u}_{is})}{\sqrt{\text{Var}(\tilde{u}_{it})\text{Var}(\tilde{u}_{is})}}$$

$$\text{With } \text{Var}(\tilde{u}_{it}) = E(\tilde{u}_{it}^2) = E(u_{it} - \bar{u}_i)^2$$

$$= E(\tilde{u}_{it}^2) + E(\bar{u}_i^2) - 2E(u_{it} \bar{u}_i)$$

$$= \sigma_u^2 + \frac{\sigma_u^2}{T} - 2\frac{\sigma_u^2}{T}$$

$$\text{Hence } \text{corr}(\tilde{u}_{it} \tilde{u}_{is}) = \frac{-\sigma_u^2/T}{\sigma_u^2 + \frac{\sigma_u^2}{T} - 2\frac{\sigma_u^2}{T}}$$

$$\text{corr}(\tilde{u}_{it} \tilde{u}_{is}) = \frac{-\sigma_u^2/T}{T\sigma_u^2 - \sigma_u^2} = \frac{-1}{T-1}$$

so within-group residuals are also serially correlated, but as  $T \rightarrow \infty$  then  $\rho \rightarrow 0$

(note that with  $T=3$  the degree of autocorrelation is the same as that with 1<sup>st</sup> differencing. When  $T>3$  then autocorrelation in the within-group model is smaller than in the 1<sup>st</sup> differencing model)

3. Given

$$y_{it} = b_1 + b_2D_t + b_3T_{it} + b_4T_t^*D_t + u_{it}$$

$$t=1,2 \\ i=1,2 \dots N$$

Then at time  $t=1$

$$Y_{i1} = b_1 \quad \text{if } T=0$$

$$Y_{i1} = b_1 + b_3 \quad \text{if } T=1$$

At time  $t=2$

$$Y_{i2} = b_1 + b_2 + b_3 + b_4 \quad \text{If } T=1 \text{ \& } D=1$$

$$Y_{i2} = b_1 + b_2 \quad \text{If } T=0 \text{ \& } D=1$$

So the change in the value of the dependent variable for those given the treatment is

$$(Y_{i2} - Y_{i1})^{\text{treated}} = (b_1 + b_2 + b_3 + b_4) - (b_1 + b_3) = b_2 + b_4$$

and for the control group

$$(Y_{i2} - Y_{i1})^{\text{control}} = (b_1 + b_2) - (b_1) = b_2$$

so the difference in difference estimator is given by

$$(Y_{i2} - Y_{i1})^{\text{treated}} - (Y_{i2} - Y_{i1})^{\text{control}} = b_4$$

which is the coefficient on the interaction term  $T^*D$  in the model above

If  $T > 2$  then there are effectively several difference in differences in existence

$$(Y_{it} - Y_{i,t-1})^{\text{treated}} - (Y_{it} - Y_{i,t-1})^{\text{control}} \quad \text{for } t=2,3..T$$

The problem then becomes whether the average of the difference-in-differences is significantly different in the period after the intervention took place compared with the average in the period before

Suppose the intervention occurs at  $t=2$

$$\text{It follows that} \quad y_{it} = b_1 + b_2Y_t + b_3T_{it} + b_4Y_t^*D_t + u_{it} \quad t=1,2,..T$$

will do this since

the d-i-d before the intervention took place is

$$(Y_{i2} - Y_{i1})^{\text{treated}} - (Y_{i2} - Y_{i1})^{\text{control}} = [(b_1 + b_{22} + b_3) - (b_1 + b_{22} + b_3)] - [(b_1 + b_{22}) - (b_1)] = 0$$

but in the period after the intervention

$$(Y_{i3} - Y_{i2})^{\text{treated}} - (Y_{i3} - Y_{i2})^{\text{control}} = [(b_1 + b_{23} + b_3 + b_4) - (b_1 + b_{22} + b_3)] - [(b_1 + b_{23}) - (b_1 + b_{22})] = b_4$$

and this holds for any  $t > 2$

So the coefficient  $b_4$  measures the average

4. Since consistency of the within-group estimator relies on

$$p \lim \left( \frac{\sum_{i=1}^N \sum_{t=1}^T (X_{it} - \bar{X}_i)(u_{it} - \bar{u}_i)}{NT} \right) = 0$$

$$\text{where } \bar{X}_i = \frac{\sum_{t=1}^T X_{it}}{T} = \frac{X_{i1} + X_{i2} + \dots + X_{iT}}{T}$$

this effectively requires that the residual at time  $t$  is uncorrelated with the  $X$  values in **ALL** time periods (this is called "strong exogeneity")

and so any model with lagged values of  $X$  will be endogenous if estimated using within-groups estimation

5. Given the between group estimator

$$(1) \quad \bar{y}_i = b_0 + b_1 \bar{x}_i + a_i + \bar{u}_i \quad i = 1, \dots, N$$

then OLS on (1)

$$\hat{\beta} = \left( \bar{X}' \bar{X} \right)^{-1} \bar{X}' \bar{y}$$

and consistency of this estimator given by 
$$p \lim \hat{\beta} = p \lim \left( \frac{\bar{X}' \bar{X}}{T} \right)^{-1} p \lim \left( \frac{\bar{X}' \bar{y}}{T} \right)$$

$$p \lim \hat{\beta} = p \lim \left( \frac{\bar{X}' \bar{X}}{T} \right)^{-1} p \lim \left( \frac{\bar{X}' (\bar{X} \beta + a_i + \bar{v}_i)}{T} \right)$$

$$p \lim \hat{\beta} = \beta + p \lim \left( \frac{\bar{X}' (a_i + \bar{v}_i)}{T} \right) p \lim \left( \frac{\bar{X}' \bar{X}}{T} \right)^{-1}$$

$$p \lim \hat{\beta} = \beta + \frac{\sigma \bar{X} a}{\sigma^2 \bar{X}}$$

where 
$$p \lim \left( \frac{\bar{X}' a_i}{T} \right) = \sigma \bar{X} a \quad \text{and} \quad p \lim \left( \frac{\bar{X}' \bar{X}}{T} \right) = \sigma^2 \bar{X}$$

$$\left( \text{Assuming } p \lim \left( \frac{\bar{X}' \bar{v}_i}{T} \right) = 0 \right)$$

Since these are grouped residuals use result from Exercise 4 (Heteroskedasticity) that

$$\sigma^2_{\bar{X}} = \text{Var}(\bar{X}) = \frac{\sigma^2_X}{T}$$

$$\text{and so } p \lim \hat{\beta} = \beta + T \frac{\sigma_{\bar{X}} a}{\sigma^2_X} \neq \beta$$

So in general the between-group estimator is inconsistent if the unobserved components are correlated with the observed variable means

and the inconsistency in the between groups estimator will **increase** with the number of time periods  $T$

(with the direction of the inconsistency depending on the covariance between  $a_i$  and the group means  $\bar{X}$  )

6. To show that pooled OLS is a weighted average of the within and between group estimators

-Assuming there are no unobserved individual effects then the pooled OLS model is

$$y_{it} = X_{it}b + u_{it} \quad (1)$$

Summing over all observations on individual  $i$  and dividing by  $T$  (the number of time series observations) gives the between group means model

$$\bar{y}_i = \bar{X}_i b + \bar{u}_i \quad (2)$$

and (1) –(2) gives the within-group estimator

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)b + (u_{it} - \bar{u}_i)$$

$$\text{where } \bar{y}_i = \frac{\sum_{t=1}^T y_{it}}{T} \quad \bar{X}_i = \frac{\sum_{t=1}^T X_{it}}{T}$$

We know (see lecture notes) that the matrices of sums and cross products can be written as

$$(X'X) = \left( \begin{array}{cc} N & T \\ \sum_{i=1}^N \sum_{t=1}^T x_{it} x_{it}' & \end{array} \right) \quad X'y = \left( \begin{array}{cc} N & T \\ \sum_{i=1}^N \sum_{t=1}^T x_{it} y_{it} & \end{array} \right) \quad y'y = \left( \begin{array}{cc} N & T \\ \sum_{i=1}^N \sum_{t=1}^T y_{it} y_{it} & \end{array} \right)$$

Similarly the within-group equivalents are

$$\left( \begin{array}{cc} N & T \\ \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' & \end{array} \right) \quad \left( \begin{array}{cc} N & T \\ \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) & \end{array} \right) \quad \left( \begin{array}{cc} N & T \\ \sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)(y_{it} - \bar{y}_i)' & \end{array} \right)$$

ie the deviations of each individual observation from its **group** mean

and the between group sums of squares are given by the deviation of each group mean from the overall sample mean (across  $N$  individuals in any one of  $T$  time periods)

$$\left( \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right) \quad \left( \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})' \right) \quad \left( \sum_{i=1}^N (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})' \right)$$

where  $\bar{y} = \frac{\sum_{i=1}^N \bar{y}_i}{N}$        $\bar{x} = \frac{\sum_{i=1}^N \bar{x}_i}{N}$

It follows that the total variation in  $X$  across  $N$  individuals and  $T$  time periods can be written as

$$\left( \sum_{t=1}^T \sum_{i=1}^N (x_{it} - \bar{x})(x_{it} - \bar{x})' \right) = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right) + \left( \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right)$$

ie Total Variation = Within-Group Variation + Between-Group Variation  
 $S_{xx} = S_{xx}^{within} + S_{xx}^{between}$

(a similar expression holds for  $Y$  or any continuous variable)

and also that

$$\left( \sum_{t=1}^T \sum_{i=1}^N (x_{it} - \bar{x})(y_{it} - \bar{y})' \right) = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)' \right) + \left( \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})' \right)$$

$$S_{xy} = S_{xy}^{within} + S_{xy}^{between}$$

It follows that the Pooled OLS estimator of  $b$  in (1) in mean deviation form can be written as

$$\hat{\beta}_{OLS} = \left( \sum_{t=1}^T \sum_{i=1}^N (x_{it} - \bar{x})(x_{it} - \bar{x})' \right)^{-1} \left( \sum_{t=1}^T \sum_{i=1}^N (x_{it} - \bar{x})(y_{it} - \bar{y})' \right) \quad (4)$$

which using the above means that  $\hat{\beta}_{OLS} = \left( S_{xx}^{within} + S_{xx}^{between} \right)^{-1} \left( S_{xy}^{within} + S_{xy}^{between} \right)$  (5)

Similarly the between group estimator in (2) can be written as

$$\hat{\beta}_{between} = \left( S_{xx}^{between} \right)^{-1} \left( S_{xy}^{between} \right) \quad (6)$$

and the within group estimator from (3)  $\hat{\beta}_{within} = \left( S_{xx}^{within} \right)^{-1} \left( S_{xy}^{within} \right)$  (7)

so from (6)  $S_{xx}^{within} \hat{\beta}_{within} = S_{xy}^{within}$

and from (7)

sub. this into (5) gives

$$\hat{\beta}_{OLS} = \left( S_{xx}^{within} + S_{xx}^{between} \right)^{-1} \left( S_{xx}^{within} \hat{\beta}_{within} + S_{xx}^{between} \hat{\beta}_{between} \right)$$

$$= \frac{S_{xx}^{within}}{\left( S_{xx}^{within} + S_{xx}^{between} \right)} \hat{\beta}_{within} + \frac{S_{xx}^{between}}{\left( S_{xx}^{within} + S_{xx}^{between} \right)} \hat{\beta}_{between}$$

so that the pooled OLS estimator is a weighted average of the within and between group estimates where the weight reflects the contribution of each component to the total variation in X. The larger (smaller) the variation in X between groups the closer (further away) the pooled estimator is to the between group estimate (and therefore more like a cross-section estimate).

7. One assumption needed for unbiased OLS estimates in any model is that

$$E(X'u) = 0$$

which in turn requires that for each row of the X matrix  $E[X_i(y_i - X_i\beta)] = 0$

This is called a "moment restriction"

and the sample equivalent to this is 
$$\frac{1}{N} \sum_{i=1}^N X_i u_i = \frac{1}{N} \sum_{i=1}^N X_i (y_i - X_i \hat{\beta})$$

(so that on average the sample correlation between the observed x values and the residuals is zero)

Note that the value of  $\beta$  which minimises this value is exactly that which minimises the sum of squares in the normal equations used to define the OLS estimator

[F.O.C. minimum

$$\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

which gives k normal equations 
$$X'X\hat{\beta} = X'y \quad ]$$

For this reason OLS is also known as a "method of moments" estimator

For the within-group model this is based on the idea that

$$E \left( X_{it} - \bar{X}_i \right) \left( \bar{u}_{it} \right) = 0 \quad \text{for all } i \text{ and all } t$$

where the within group residual 
$$\bar{u}_{it} = (Y_{it} - \bar{Y}_i) - \left( X_{it} - \bar{X}_i \right) \hat{\beta} = 0$$

The moment condition (equivalent to minimising the sum of squares) can therefore be written as

$$\sum_{i=1}^N \sum_{t=1}^2 \left( X_{it} - \bar{X}_i \right) \left( Y_{it} - \bar{Y}_i - \left( X_{it} - \bar{X}_i \right) \hat{\beta} \right) = 0 \quad (1)$$

(this is called the sample moment restriction since it is unique to this particular sample)

when  $T=2$  then

$$\bar{X}_i = \frac{\sum_{t=1}^2 X_{it}}{2} = \frac{X_{i1} + X_{i2}}{2} \quad \bar{Y}_i = \frac{\sum_{t=1}^2 Y_{it}}{2} = \frac{Y_{i1} + Y_{i2}}{2}$$

and (1) can be written as

$$\begin{aligned} & \sum_{i=1}^N \left( X_{i1} - \left( \frac{X_{i1} + X_{i2}}{2} \right) \right) \left( Y_{i1} - \left( \frac{Y_{i1} + Y_{i2}}{2} \right) - \left( X_{i1} - \left( \frac{X_{i1} + X_{i2}}{2} \right) \right) \hat{\beta} \right) \\ & + \sum_{i=1}^N \left( X_{i2} - \left( \frac{X_{i1} + X_{i2}}{2} \right) \right) \left( Y_{i2} - \left( \frac{Y_{i1} + Y_{i2}}{2} \right) - \left( X_{i2} - \left( \frac{X_{i1} + X_{i2}}{2} \right) \right) \hat{\beta} \right) = 0 \\ & = \sum_{i=1}^N \left( \frac{X_{i1} - X_{i2}}{2} \right) \left( \left( \frac{Y_{i1} - Y_{i2}}{2} \right) - \left( \left( \frac{X_{i1} - X_{i2}}{2} \right) \hat{\beta} \right) \right) \\ & \quad + \sum_{i=1}^N \left( \frac{X_{i2} - X_{i1}}{2} \right) \left( \left( \frac{Y_{i2} - Y_{i1}}{2} \right) - \left( \left( \frac{X_{i2} - X_{i1}}{2} \right) \hat{\beta} \right) \right) = 0 \\ & = \sum_{i=1}^N - \left( \frac{X_{i2} - X_{i1}}{2} \right) \left( - \left( \frac{Y_{i2} - Y_{i1}}{2} \right) - \left( - \left( \frac{X_{i2} - X_{i1}}{2} \right) \hat{\beta} \right) \right) \\ & \quad + \sum_{i=1}^N \left( \frac{X_{i2} - X_{i1}}{2} \right) \left( \left( \frac{Y_{i2} - Y_{i1}}{2} \right) - \left( \left( \frac{X_{i2} - X_{i1}}{2} \right) \hat{\beta} \right) \right) = 0 \\ & \Rightarrow \sum_{i=1}^N (X_{i2} - X_{i1}) \left( (Y_{i2} - Y_{i1}) - (X_{i2} - X_{i1}) \hat{\beta} \right) = 0 \end{aligned}$$

which is exactly the 1<sup>st</sup> order condition used to minimise the sum of squares in the 1<sup>st</sup> differenced model

$$\Delta y = \Delta X \beta + \Delta u$$

$$Y_{i2} - Y_{i1} = (X_{i2} - X_{i1}) \beta + u_{i2} - u_{i1}$$

Note that this equivalence does not hold when  $T > 2$  since

$$\sum_{i=1}^N \sum_{t=1}^3 \left( X_{it} - \bar{X}_i \right) \left( Y_{it} - \bar{Y}_i \right) - \left( X_{it} - \bar{X}_i \right) \hat{\beta} \neq \sum_{i=1}^N \sum_{t=2}^3 \left( X_{it} - X_{i,t-1} \right) \left( y_{it} - y_{i,t-1} \right) - \left( X_{it} - X_{i,t-1} \right) \hat{\beta}$$

8. Show that the fixed effects estimator is consistent in an unbalanced panel

In an unbalanced panel some observations for some (but not necessarily all) individuals are missing in certain time periods

Given the panel data model  $y_{it} = X_{it}b + u_{it}$

The usual within-group fixed effects estimator can be written as

$$\hat{\beta}_{FE} = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right) \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \right) = \left( \sum_{i=1}^N \sum_{t=1}^T x_{it} x_{it}' \right) \left( \sum_{i=1}^N \sum_{t=1}^T x_{it} y_{it} \right)$$

Let  $d_i = (d_{i1}, d_{i2}, \dots, d_{iT})$  be a 1 by  $T$  vector of dummy variables (sometimes called selection indicators) taking the value 1 if individual  $i$  is observed in time period  $t$ , equal 0 otherwise. Then for an unbalanced

panel the fixed effect estimator can be written as  $\hat{\beta}_{fe} = \left( \sum_{i=1}^N \sum_{t=1}^T s_{it} x_{it} x_{it}' \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T s_{it} x_{it} y_{it} \right)$

$$= \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T s_{it} x_{it} x_{it}' \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T s_{it} x_{it} y_{it} \right)$$

Sub. in for  $y_{it}$

$$\hat{\beta}_{FE} = \hat{\beta} + \left( \frac{\sum_{i=1}^N \sum_{t=1}^T s_{it} x_{it} x_{it}'}{N} \right)^{-1} \left( \frac{\sum_{i=1}^N \sum_{t=1}^T s_{it} x_{it} u_{it}}{N} \right) \quad (1)$$

$$\text{So consistency depends on } \left( \frac{\sum_{i=1}^N \sum_{t=1}^T s_{it} x_{it} u_{it}}{N} \right) = \left( \frac{\sum_{i=1}^N \sum_{t=1}^T s_{it} \left( x_{it} - \bar{x}_i \right) u_{it}}{N} \right)$$

Since this involves the within-group mean of  $x$  this relies on a strict exogeneity condition – the assumption that the residual in period  $t$  is uncorrelated with the  $x$  values from **any** period (just as the consistency requirement for a balanced panel). The introduction of the indicator term  $s_i$  does not change this requirement and so the proof of consistency is the same

(note that the convergence to a finite value of the first term in (1) in the limit rules out any time-invariant variables, since they would be collinear and so an inverse of this matrix will not exist)

9. Consider a simple 2 variable panel data model

$$y_{it} = bx_{it}^* + f_i + u_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T$$

where now the x variable is measured with error  $x_{it}^* = x_{it} + e_{it}$

$$\text{so that the observed model is } y_{it} = bx_{it} + f_i + (u_{it} + be_{it}) = bx_{it} + f_i + v_{it} \quad (2)$$

Consider the consistency of the pooled OLS estimator

$$\text{OLS on (2) gives } \hat{\beta} = (x'x)^{-1}x'y = \left(\frac{x'x}{NT}\right)^{-1} \left(\frac{x'y}{NT}\right)$$

$$\text{Consider } p \lim \left(\frac{x'y}{NT}\right) = p \lim \left( \frac{\sum_{i=1}^N \sum_{t=1}^T x_{it} y_{it}}{NT} \right) = p \lim \left( \frac{\sum_{i=1}^N \sum_{t=1}^T (x_{it}^* - e_{it})(bx_{it}^* + f_i + v_{it})}{NT} \right)$$

$$= b\sigma_x^2 + \sigma_x^* f \quad \text{since all other correlations vanish at the limit}$$

Similarly

$$p \lim \left(\frac{x'x}{NT}\right) = p \lim \left( \frac{\sum_{i=1}^N \sum_{t=1}^T x_{it} x_{it}'}{NT} \right) = p \lim \left( \frac{\sum_{i=1}^N \sum_{t=1}^T (x_{it}^* - e_{it})(x_{it}^* - e_{it})}{NT} \right)$$

$$= \sigma_x^2 + \sigma_e^2$$

$$\text{So } p \lim \hat{\beta}_{OLS} = \frac{b\sigma_x^2 + \sigma_x^* f}{\sigma_x^2 + \sigma_e^2} = b + \frac{\sigma_x^* f}{\sigma_x^2 + \sigma_e^2} - \frac{b\sigma_e^2}{\sigma_x^2 + \sigma_e^2} \quad (3)$$

Hence in the presence of measurement error in panel data the pooled OLS estimator is inconsistent due to

- the usual attenuation effect from measurement error (the third term in (3))
- an additional bias due to the presence of the fixed effect  $f_i$

*(note that the fixed effects estimator will be inconsistent only because of any measurement error. A between-group estimator in the presence of measurement error can be shown to be consistent with the individual specific dummies acting as instruments for the mis-measured x variable)*