Panel data

1 Panel data

Cross-sectional data typically collect information on N units (individuals, firms, countries, etc.) at a given moment in time. Time-series data collect information on a single unit (typically, a country) over many (T) time periods. So far, we have concentrated our attention on cross-sectional data and ignored time-series data. Both types of data move along a single dimension: individuals (cross-sectional data) or time (time-series data).

Panel data (or longitudinal data) move instead along two (or more) dimensions. The most frequent case is one where N units are observed over T time periods. The total number of observations is therefore NT. Three dimensions of data are also possible: for example, we could have information on N multinational firms followed over T time periods and over the C countries where they operate. The total number of observations in this case is NTC.

Two-dimension panels do not necessarily extend over units and time periods. We could have two-dimension panels extending over units and sub-units, such as when we have data on siblings within families, or branches within firms, or counties within states.

Panels can be *balanced* (when the units are followed for an equal number of periods) or *unbalanced* (when the units are followed for a different number of periods). Panels can be *continuous* (when there is no refreshing, i.e., no entry of new units -an example is the Panel Study of Income Dynamics) or *rotating* (when units leave and are replaced by new units -an example is the Consumer Expenditure Survey).

1.1 Advantages of panel data

- Panel data allow us to control for unobserved heterogeneity. For example, in the classical Mundlak's example, we can control directly for managerial ability and inputs use, so as to eliminate the omitted variable bias. In general, any unobserved heterogeneity component that remains fixed over time can be handled thus reducing considerably the omitted variable bias problem.

- They offer obvious statistical advantages. They may help us reduce the problem of collinearity among variables, and may give us more precise estimates due to the efficiency gain brought by more data.

- They can help us addressing questions of dynamics as discussed in class.

1.2 Disadvantages of panel data

- Panel data suffer from attrition, i.e., the fact that people who are initially part of the data set fail to be interviewed in subsequent waves. For example, people may move and the survey may be unable to track them down. The attrition of units would not be problematic if attrition were completely at random. However, this is unlikely to be the case. If we are studying the dynamics of poverty, for example, we may be worried that the people who move (and therefore disappear from the survey) are the poorest, for example due to lack of job opportunities in the local labor market or, to the extreme, because they become homeless. Since the poor disappears, we will have the impression that things are improving simply because the only people who remain in the panel are the rich.

1.3 Prototypical panel data model

The prototypical panel data model is

$$y_{it} = \alpha + \gamma x_{it} + f_i + v_{it}$$

for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$ for a total of NT observations. This is a model with a single regressor x_{it}. The new term is f_i, which is assumed to be individual-specific and time invariant. The classical reference is Mundlak, who though of y_{it} as the output of a farm, x_{it} as inputs (fertilizers, etc.), f_i as managerial ability, and v_{it} as random event outside the manager's control (such as weather).

Here we will make the assumption that $E(v_{it}|f_i, x_{it}) = 0$. As for f_i, two models are usually proposed based on the assumption one is willing to make regrading the correlation between f and x:

(a) the random effect model, corresponding to $E(f_i|x_{it}) = 0$;

(b) the fixed effect model, corresponding to $E(f_i|x_{it}) \neq 0$.

In case (a), we assume that f_i is an error component, while in case (b) we assume it is a parameter to estimate. In the example above, if we believe that managers' ability is randomly allocated among farms, the random effect is a good description of the reality; in contrast, if we believe that managers' ability is correlated with inputs choice and efficiency, the fixed effect model should be used. The main difference between the two models is that they lead to two different estimators.

2 Random effect model

(Wooldridge, ch. 14.2)

The random effect model is a variant of GLS (generalized least squares). GLS is used when the error term of the regression

$$y = X\beta + u$$

is heteroskedastic, i.e., $E(uu') \neq \sigma^2 I$.

Let's take the prototypical panel data model of the previous section:

$$\begin{aligned} y_{it} &= \alpha + \gamma x_{it} + \underbrace{(f_i + v_{it})}_{u_{it}=\text{error term}} \\ &= \alpha + \gamma x_{it} + u_{it} \end{aligned}$$

for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$. In matrix format

$$\underset{(NT\times 1)}{\mathbf{y}} = \underset{(NT\times 2)}{\mathbf{X}} \underset{(2\times 1)}{\beta} + \underset{(NT\times 1)}{\mathbf{u}}$$

Let's assume that f_i is a zero-mean error term with variance σ_f^2 and u_{it} is an i.i.d. zero-mean error term with variance σ_v^2. As well, f_i and v_{it} are orthogonal ($E(f_i v_{is}) = 0$ for all s). The "global" error term u_{it} is indeed heteroskedastic. In fact:

$$\begin{aligned} E\left(u_{it}^2\right) &= E\left((f_i + v_{it})^2\right) = E\left(f_i^2\right) + E\left(v_{it}^2\right) + E\left(f_i v_{it}\right) \\ &= \sigma_f^2 + \sigma_v^2 \end{aligned}$$

and for $t \neq s$

$$\begin{aligned} E\left(u_{it} u_{is}\right) &= E\left((f_i + v_{it})(f_i + v_{is})\right) = E\left(f_i^2\right) + E\left(f_i v_{is}\right) + E\left(f_i v_{it}\right) + E\left(v_{it} v_{is}\right) \\ &= \sigma_f^2 \end{aligned}$$

Thus, if $\mathbf{u}_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ ... \\ u_{it} \end{pmatrix}$ for a generic individual i, we have that his/her variance matrix of the error is:

$$\begin{aligned} \underset{(T\times T)}{E\left(\mathbf{u}_i \mathbf{u}_i'\right)} &= \begin{pmatrix} \sigma_f^2 + \sigma_v^2 & \sigma_f^2 & ... & \sigma_f^2 \\ \sigma_f^2 & \sigma_f^2 + \sigma_v^2 & ... & \sigma_f^2 \\ ... & ... & ... & ... \\ \sigma_f^2 & \sigma_f^2 & ... & \sigma_f^2 + \sigma_v^2 \end{pmatrix} \\ &= \sigma_v^2 \mathbf{I} + \sigma_f^2 \mathbf{i}\mathbf{i}' \\ &= \mathbf{\Omega} \end{aligned}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{i}$ a vector of ones. Stacking the vectors $\mathbf{u}_i$ for all individuals, we get that the error term of the pooled regression is: $\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ ... \\ \mathbf{u}_N \end{pmatrix}$, and so its variance matrix

$$\underset{(NT \times NT)}{E(\mathbf{uu}')} = \begin{pmatrix} \mathbf{\Omega} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega} & ... & \mathbf{0} \\ ... & ... & ... & ... \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{\Omega} \end{pmatrix} = \mathbf{\Omega} \otimes \mathbf{I}$$

is a block-diagonal matrix, and $\otimes$ indicates a Kronecker product.

2.1 GLS

The general model

$$y = X\beta + u \tag{1}$$

could certainly be estimated by OLS even in the presence of heteroskedasticity of the error term. The estimates will be unbiased and consistent. To see this, suppose for simplicity that X is fixed (non-stochastic). Then OLS is

$$\begin{aligned} \hat{\beta}_{OLS} &= (X'X)^{-1} X'y \\ &= \beta + (X'X)^{-1} X'u \end{aligned}$$

Its expectation is

$$\begin{aligned} E\left(\hat{\beta}_{OLS}\right) &= \beta + (X'X)^{-1} X'E(u) \\ &= \beta \end{aligned}$$

because $E(u) = 0$. It is also easy to prove consistency of $\hat{\beta}_{OLS}$.

What about efficiency? Assume $E(uu') = \sigma^2\Sigma$, so that there's heteroskedasticity. The "true" variance of $\hat{\beta}_{OLS}$ (i.e., the variance matrix that should be used to conduct inference) is

$$\begin{aligned} var\left(\hat{\beta}_{OLS}\right)_{HT} &= var\left(\beta + (X'X)^{-1} X'u\right) \\ &= (X'X)^{-1} X'var(u) X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} X'\Sigma X (X'X)^{-1} \end{aligned}$$

where $var\left(\hat{\beta}_{OLS}\right)_{HT}$ means that this is the variance of the OLS estimator when the econometician knows that there is heteroskedasticity.

There are two important things to notice here:

(a) OLS is no longer BLUE. We can find another estimator that is also unbiased, linear, but is *more* efficient than OLS. This estimator is the GLS estimator (see below).

(b) An econometrician that estimates the model (1) by OLS ignoring heteroskedasticity of the error term will conduct inference on β using the wrong variance matrix of the estimates, i.e., he/she will use $var\left(\hat{\beta}_{OLS}\right)_{HM} = \sigma^2 \left(X'X\right)^{-1}$ instead of the correct $var\left(\hat{\beta}_{OLS}\right)_{HT}$ (given above). Thus hypothesis tests will give wrong results.

How do we get the GLS estimator? The idea is to transform a heteroskedastic error term into an homoskedastic error term, and then apply the OLS idea to the transformed model.

Suppose you can find a (non-singular, square, and non-stochastic) matrix P such that

$$\Sigma^{-1} = P'P$$

(see Johnston's chapter for more details on this). Then take the heteroskedastic model

$$y = X\beta + u$$

with $E\left(uu'\right) = \sigma^2\Sigma$. Pre-multiply both sides by P:

$$\begin{aligned} Py &= PX\beta + Pu \\ \tilde{y} &= \tilde{X}\beta + \tilde{u} \end{aligned}$$

Is $\tilde{u}$ homoskedastic? First note that $E\left(\tilde{u}\right) = E\left(Pu\right) = PE\left(u\right) = 0$. Then note that

$$var\left(\tilde{u}\right) = var\left(Pu\right) = Pvar\left(u\right)P' = \sigma^2 P\Sigma P'$$

Now we will prove that $P\Sigma P' = I$ starting from the definition

$$\begin{aligned} \Sigma^{-1} &= P'P \\ \Sigma^{-1}P^{-1} &= P'PP^{-1} \\ \left(P'\right)^{-1}\Sigma^{-1}P^{-1} &= \underbrace{\left(P'\right)^{-1}P'}_{I}\underbrace{PP^{-1}}_{I} \\ \left(P'\right)^{-1}\Sigma^{-1}P^{-1} &= I \end{aligned}$$

and since $\left(ABC\right)^{-1} = C^{-1}B^{-1}A^{-1}$, it follows that

$$P\Sigma P' = I$$

and therefore

$$var\left(\tilde{u}\right) = \sigma^2 P\Sigma P' = \sigma^2 I$$

is homoskedastic. Thus OLS applied to the transformed model will give unbiased and consistent estimates of β and will also be BLUE:

$$\begin{aligned}
\hat{\beta}_{OLS,tm} &= \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y} \\
&= \left(X'P'PX\right)^{-1}X'P'Py \\
&= \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}y \\
&= \hat{\beta}_{GLS}
\end{aligned}$$

where $\hat{\beta}_{OLS,tm}$ stands for OLS applied to the transformed model. The variance of this estimator, still assuming X is fixed for simplicity, is

$$\begin{aligned}
var\left(\hat{\beta}_{GLS}\right) &= var\left(\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}y\right) \\
&= var\left(\beta + \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}u\right) \\
&= \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}var\left(u\right)\Sigma^{-1}X\left(X'\Sigma^{-1}X\right)^{-1} \\
&= \sigma^2\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}\Sigma\Sigma^{-1}X\left(X'\Sigma^{-1}X\right)^{-1} \\
&= \sigma^2\left(X'\Sigma^{-1}X\right)^{-1}
\end{aligned}$$

This estimator can also be obtained by

$$\underset{\hat{\beta}}{min}\hat{u}'\Sigma^{-1}\hat{u}$$

so it's similar to the IV idea of weighting residuals differently instead of equally (as in OLS). The solution of this problem is indeed

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

Intuitively, the GLS estimator puts more weight on observations with a lower variance of the error. That's where its precision gain comes from relative to OLS. The general form of the GLS estimator in a model

$$y = X\beta + u$$

with u having a generic variance matrix $var\left(u\right)$ is:

$$\hat{\beta}_{GLS} = \left(X'var\left(u\right)^{-1}X\right)^{-1}X'var\left(u\right)^{-1}y$$

Hence, for the random effect model in which $var\left(u\right) = \Omega \otimes I$ we have

$$\hat{\beta}_{RE} = \left(X'\left(\Omega \otimes I\right)^{-1}X\right)^{-1}X'\left(\Omega \otimes I\right)^{-1}y$$

2.2 Feasible GLS

There is one obvious problem in the implementation of the random effect estimator (and in general in the implementation of any GLS estimator): $\hat{\beta}_{RE}$ depends on Ω, which is unknown unless one has some outside information about σ_f^2 and σ_v^2 (an unlikely case). One way out is to replace Ω with a consistent estimate of it, say $\hat{\Omega}$. If this is done, one implements a so called feasible RE (or more generally, feasible GLS) estimator:

$$\hat{\beta}_{FRE} = \left(X'\left(\widehat{\Omega} \otimes I\right)^{-1} X\right)^{-1} X'\left(\widehat{\Omega} \otimes I\right)^{-1} y$$

To obtain an estimate of $\widehat{\Omega}$,we need to find estimates of the two parameters that appear in it, σ_f^2 and σ_v^2. We will use a couple of results:

(1) Suppose you estimate the model by OLS. As said, OLS gives consistent estimates. Define the OLS residual $\widehat{u}_{it} = y_{it} - \hat{\alpha}_{OLS} - \hat{\gamma}_{OLS} x_{it}$. In general, we have learned that an estimate of the variance of the error term when we have a cross-section of N individuals and a model with k regressors is

$$\frac{\sum_{i=1}^{N} \hat{u}_{it}^2}{N-k}$$

We proved that $E\left(\frac{\sum_{i=1}^{N} \hat{u}_{it}^2}{N-k}\right) = \sigma_u^2$. In panel data, however, we have a total of NT observations, so that the analog for panel data is

$$\widehat{\sigma_u^2} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2}{NT-k}$$

which is also an unbiased estimate of σ_u^2.

(2) Since $u_{it} = f_i + v_{it}$ and f_i and v_{it} are both mean zero and orthogonal to each other,

$$\sigma_u^2 = E\left(u_{it}^2\right) = \sigma_f^2 + \sigma_v^2$$

Thus if we have an estimate of σ_u^2 (calculated using result (1) above) and an estimate of one of its two component (σ_f^2 or σ_v^2), the other one can be determined by difference.

Now we have to find an estimate of σ_f^2 (which is easier than finding an estimate of σ_v^2). Note that

$$E\left(u_{is} u_{it}\right) = \sigma_f^2 \text{ for all } s \neq t$$

Consider a generic individual i. There are $\frac{(T-1)T}{2}$ possible $u_{is} u_{it}$ products, namely $u_{i1}u_{i2}$,$u_{i1}u_{i3}$, ..., $u_{i1}u_{iT}$,$u_{i2}u_{i3}$,...,$u_{iT-1}u_{iT}$. Taking their sum:

$$\sum_{s=1}^{T-1} \sum_{t=s+1}^{T} u_{is} u_{it}$$

the expectation of which is

$$E\left(\sum_{s=1}^{T-1}\sum_{t=s+1}^{T}u_{is}u_{it}\right)=\sum_{s=1}^{T-1}\sum_{t=s+1}^{T}E\left(u_{is}u_{it}\right)=\sum_{s=1}^{T-1}\sum_{t=s+1}^{T}\sigma_f^2=\frac{(T-1)\,T}{2}\sigma_f^2$$

and hence $E\left(\frac{\sum_{s=1}^{T-1}\sum_{t=s+1}^{T}u_{is}u_{it}}{\frac{(T-1)T}{2}}\right)=E\left(\frac{\sum_{i=1}^{N}\sum_{s=1}^{T-1}\sum_{t=s+1}^{T}u_{is}u_{it}}{N\frac{(T-1)T}{2}}\right)=\sigma_f^2$, where the second equality comes from the fact that we can use any individual in the sample.

We don't have the true errors u, just the residuals $\hat{u}$. We can use these as long as we make a degrees of freedom correction and we obtain:

$$E\left(\underbrace{\frac{\sum_{s=1}^{T-1}\sum_{t=s+1}^{T}\hat{u}_{is}\hat{u}_{it}}{N\frac{(T-1)T}{2}-k}}_{\widehat{\sigma_f^2}}\right)=\sigma_f^2$$

At this point we can obtain:

$$\widehat{\sigma_v^2}=\widehat{\sigma_u^2}-\widehat{\sigma_f^2}$$

and construct $\hat{\Omega}$ and implement FRE (or FGLC in general).

To summarize, FGLS is obtained using a two step procedure:

(a) Run OLS on the heteroskedastic model $y=X\beta+u$, and construct the residual $\hat{u}=y-X\hat{\beta}_{OLS}$. Use this to obtain estimates $\widehat{\sigma_u^2}$ and $\widehat{\sigma_f^2}$ and thus by difference $\widehat{\sigma_v^2}$. This allows to construct $\hat{\Omega}$.

(b) The FGLS estimator is $\hat{\beta}_{FGLS}=\left(X'\left(\widehat{\Omega}\otimes I\right)^{-1}X\right)^{-1}X'\left(\widehat{\Omega}\otimes I\right)^{-1}y$ in the random effect model case. Use the correct variance matrix of $\hat{\beta}_{FGLS}$, $var\left(\hat{\beta}_{FGLS}\right)=\left(X'\left(\widehat{\Omega}\otimes I\right)^{-1}X\right)^{-1}$ to conduct inference on β.

3 Fixed effect model

Recall the prototypical panel data model:

$$y_{it}=\alpha+\gamma x_{it}+f_i+v_{it}$$

for $i=1,2,...,N$ and $t=1,2,...,T$. In the fixed effect model we make the assumption $E\left(f_i|x_{it}\right)\neq 0$. The term f_i is thus interpreted as just another parameter to estimate. In practice, a dummy for individual "1" identifies f_1, a dummy for individual "2" identifies f_2, and so on. If there is an overall constant

(α), one of the fixed effects cannot be identified, or some restrictions must be imposed.

Let's assume $\alpha = 0$ so this identification problem is set aside. In matrix format:

$$y = X\beta + Df + v$$

where D is a $(NT \times N)$ matrix of dummies for each of the N individuals. This model can now be estimated by OLS because v is orthogonal to both X and D. Using results from partitioned regression, we can now write

$$\begin{aligned}
\hat{\beta} &= \left(X' M_D X\right)^{-1} X' M_D y \\
\hat{f} &= \left(D' M_X D\right)^{-1} D' M_X y
\end{aligned}$$

where $M_X = I - X\left(X'X\right)^{-1} X'$ and $M_D = I - D\left(D'D\right)^{-1} D'$ are idempotent matrices. The estimators $\hat{\beta}$ and $\hat{f}$ are known as LSDV (least squares dummy variables) estimators. They are also known as within-group estimators for reasons that will become clear soon.

$\hat{\beta}$ can actually be rewritten in a much nicer way. Suppose $N = 2$ for simplicity. Then

$$D = \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}$$

where i is a $(T \times 1)$ vector of ones. Then

$$\begin{aligned}
M_D &= I - D\left(D'D\right)^{-1} D' \\
&= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix} \left[\begin{pmatrix} i' & 0 \\ 0 & i' \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}\right]^{-1} \begin{pmatrix} i' & 0 \\ 0 & i' \end{pmatrix} \\
&= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & T \end{pmatrix}^{-1} \begin{pmatrix} i' & 0 \\ 0 & i' \end{pmatrix} \\
&= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix} \begin{pmatrix} \frac{1}{T} & 0 \\ 0 & \frac{1}{T} \end{pmatrix} \begin{pmatrix} i' & 0 \\ 0 & i' \end{pmatrix} \\
&= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix} \begin{pmatrix} \frac{1}{T}i' & 0 \\ 0 & \frac{1}{T}i' \end{pmatrix} \\
&= \begin{pmatrix} I - \frac{1}{T}ii' & 0 \\ 0 & I - \frac{1}{T}ii' \end{pmatrix}
\end{aligned}$$

What is $M_D y$ then?

$$
\begin{aligned}
M_D y &= \begin{pmatrix} I - \frac{1}{T} ii' & 0 \\ 0 & I - \frac{1}{T} ii' \end{pmatrix} y = \\
&= \begin{pmatrix} I - \frac{1}{T} ii' & 0 \\ 0 & I - \frac{1}{T} ii' \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= \begin{pmatrix} y_{11} - \frac{\sum_{t=1}^{T} y_{1t}}{T} \\ y_{12} - \frac{\sum_{t=1}^{T} y_{1t}}{T} \\ \dots \\ y_{1T} - \frac{\sum_{t=1}^{T} y_{1t}}{T} \\ y_{21} - \frac{\sum_{t=1}^{T} y_{2t}}{T} \\ y_{22} - \frac{\sum_{t=1}^{T} y_{2t}}{T} \\ \dots \\ y_{2T} - \frac{\sum_{t=1}^{T} y_{2t}}{T} \end{pmatrix}
\end{aligned}
$$

In other words, M_D is a matrix that tranforms the vector that pre-multiplies in deviation from the (individual) mean. Thus

$$
\hat{\beta} = \left(X' M_D X\right)^{-1} X' M_D y \tag{2}
$$

for the univariate case can be written as

$$
\hat{\beta} = \frac{cov\left(x_{it} - \overline{x_i}, y_{it} - \overline{y_i}\right)}{var\left(x_{it} - \overline{x_i}\right)}
$$

where $\overline{a_i}$ is the average of variable a for individual i. That's the reason why the LSDV estimator is also known as within group estimator: the variability that is used to dientify $\hat{\beta}$ is the variability that exists *within* a group (where a group here is an indivividual). The "within" part here comes from the fact that $x_{it} - \overline{x_i}$ and $y_{it} - \overline{y_i}$ would be identically zero if x_{it} and y_{it} would not take different values for a given individual over time. As a consequence, the within group estimator is not able to identify the effect of time-invariant characteristics (such as schooling).

The "within-group" interpretation just derived should make sense intuitively. Consider the prototypical model again:

$$y_{it} = f_i + \gamma x_{it} + v_{it} \tag{3}$$

Take sums of equation (3) over the T time periods the individual is observed and divide by T to obtain

$$\frac{\sum_{t=1}^{T} y_{it}}{T} = f_i + \gamma \frac{\sum_{t=1}^{T} x_{it}}{T} + \frac{\sum_{t=1}^{T} v_{it}}{T} \tag{4}$$

Now subtract (4) from (3) to obtain:

$$(y_{it} - \overline{y_i}) = \gamma (x_{it} - \overline{x_i}) + (v_{it} - \overline{v_i})$$

In this model the fixed effect no longer appears. Since x_{it} is exogenous with respect to v_{it} this equation can be estimated by OLS with no bias. The resulting OLS estimator is, indeed, the LSDV estimator (2).

As for the fixed effect, one can easily prove that the expression $\hat{f} = (D'M_X D)^{-1} D'M_X y$ can be rewritten for a generic individual i as

$$\hat{f}_i = \overline{y_i} - \overline{X}_i \hat{\beta}$$

where $\overline{X}_i$ is a row vector of means on the X for individual i (i.e., if the model is $y_{it} = f_i + \sum_{j=1}^{k} x_{jit}\beta_j + v_{it}$, then $\hat{f}_i = \overline{y_i} - \sum_{j=1}^{k} \overline{x}_{ji}\hat{\beta}_j$).

4 Inference

From partitioned regression,

$$var\left(\hat{\beta}\right) = \sigma_v^2 \left(X'M_D X\right)^{-1}$$

The error of the within group regression is $(v_{it} - \bar{v}_i)$. Note that

$$\sum_{t=1}^{T} (v_{it} - \bar{v}_i)^2 = \sum_{t=1}^{T} v_{it}^2 - T\bar{v}_i^2$$

and taking expectations

$$
\begin{aligned}
E\left(\sum_{t=1}^{T}(v_{it}-\bar{v}_i)^2\right) &= \sum_{t=1}^{T}E\left(v_{it}^2\right)-TE\left(\frac{\sum_{t=1}^{T}v_{it}}{T}\right)^2 \\
&= T\sigma_v^2 - T\frac{1}{T^2}E\left(v_{i1}^2+v_{i2}^2+...+v_{iT}^2+\text{cross-products}\right) \\
&= T\sigma_v^2-\frac{1}{T}T\sigma_v^2 \\
&= (T-1)\,\sigma_v^2
\end{aligned}
$$

or $E\left(\frac{\sum_{t=1}^{T}(v_{it}-\bar{v}_i)^2}{T-1}\right)=\sigma_v^2$.

Given that there are N individuals, and that using residuals instead of true errors require a degrees of freedom adjustment equal to the number of regressors k we have:

$$
E\left(\underbrace{\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{v_{it}-\bar{v}_i}\right)^2}{N(T-1)-k}}_{\hat{\sigma}_v^2}\right)=\sigma_v^2
$$

where $\widehat{v_{it}-\bar{v}_i}$ is the within group residual. $\hat{\sigma}_v^2$ is thus an unbiased estimate of σ_v^2 and can be used to construct an estimate of $var\left(\hat{\beta}\right)=\sigma_v^2\,(X'M_DX)^{-1}$, namely $\widehat{var\left(\hat{\beta}\right)}=\widehat{\sigma}_v^2\,(X'M_DX)^{-1}$. Note that if you do not use a built-in within group routine, but run a simple OLS regression of $(y_{it}-\overline{y_i})$ onto $(x_{it}-\overline{x_i})$, the standard errors will be incorrect, because they will use as an estimate of σ_v^2 the quantity $\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{v_{it}-\bar{v}_i}\right)^2}{NT-k}$ instead of the correct $\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{v_{it}-\bar{v}_i}\right)^2}{N(T-1)-k}$, i.e. there will be inflated standard errors.

The main problem of the Fixed Effect model is that the effect of time-invariant characteristics (such as schooling or race or gender, for example) cannot be identified (separately from the fixed effects themselves). The reason is that all the effects that are fixed over time are sweeped out by the de-meaning strategy. The problem is one of perfect collinearity. To see this, suppose you have two individuals $i=1,2$ and two time periods $t=1,2$ for simplicity. Individual i has 12 years of schooling, and individual 2 has 16. Suppose that the model we want to estimate is

$$y_{it} = f_i + \delta s_i + \gamma x_{it} + v_{it}$$

In long form

$$\begin{aligned}
\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{pmatrix} &= \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} f_1 + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} f_2 + \begin{pmatrix} 12 \\ 12 \\ 16 \\ 16 \end{pmatrix} \delta + \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{pmatrix} \gamma + \begin{pmatrix} v_{11} \\ v_{12} \\ v_{21} \\ v_{22} \end{pmatrix} \\
y &= d_1 f_1 + d_2 f_2 + s\delta + x\gamma + v
\end{aligned}$$

It is clear that $12d_1 + 16d_2 = s$, so there is perfect collinearity between d_1, d_2 and s. Thus

$$\begin{aligned}
y &= d_1 f_1 + d_2 f_2 + (12d_1 + 16d_2)\,\delta + x\gamma + v \\
&= d_1 (f_1 + 12\delta) + d_2 (f_2 + 16\delta) + x\gamma + v
\end{aligned}$$

which makes clear that the "fixed effect" that is eventually identified is a mixture of the truly unobserved fixed effect f and of all the other effects of variables that don't change over time.

Another way of eliminating the fixed effect is to take first differences instead of demeaning. Consider the prototypical model again:

$$y_{it} = f_i + \gamma x_{it} + v_{it} \tag{5}$$

and lag it one period to obtain:

$$y_{it-1} = f_i + \gamma x_{it-1} + v_{it-1} \tag{6}$$

Now subtract (6) from (5) to obtain:

$$(y_{it} - y_{t-1}) = \gamma (x_{it} - x_{it-1}) + (v_{it} - v_{it-1}) \tag{7}$$

In this model as well the fixed effect has appeared. Since x_{it} is exogenous with respect to v_{it} this equation can be estimated by OLS with no bias. Numerically, the first-difference estimate of γ will differ from the within-group estimate (unless $T = 2$ where they coincide). Asymptotically there is no difference because under the maintained assumptions they converge to the same parameter γ. However, notice that OLS applied to () as usual gives consistent and unbiased estimates, but the standard errors will be incorrect because the error term is heteroskedastic with variance matrix

$$\Omega = \begin{pmatrix} 2\sigma_v^2 & -\sigma_v^2 & ... & 0 \\ & 2\sigma_v^2 & ... & 0 \\ & & ... & ... \\ & & & 2\sigma_v^2 \end{pmatrix}$$

i.e., both the main diagonal and the one adjacent to the main one have non-zero elements. A variant of GLS (or a built-in routine) ensures efficiency and the correct standard errors.

5 Hausman test

The difference between the fixed effect model and the random effect model is the different assumption they make on $E\left(f_i|x_{it}\right)$. The RE model assumes $E\left(f_i|x_{it}\right) = 0$; the FE model that $E\left(f_i|x_{it}\right) \neq 0$. Which one is the correct one?

We can test the null that $E\left(f_i|x_{it}\right) = 0$ using a variant of the Hausman test. Under the null hypothesis $E\left(f_i|x_{it}\right) = 0$, the RE model is consistent for β and it's the most efficient estimator around (because it's BLUE). Under the same null hypothesis, FE is also consistent for β but is of course less efficient than the RE estimator. Asymptotically, the FE and the RE estimator converge to the same parameter β, and so their difference converges to zero. In small samples, they should differ only because of sampling variability if the null hypothesis is true.

Under the alternative hypothesis, $E\left(f_i|x_{it}\right) \neq 0$ and only the FE estimator remains consistent. This suggests to use the statistic

$$H = \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)' \left[var\left(\hat{\beta}_{FE}\right) - var\left(\hat{\beta}_{RE}\right)\right]^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \sim \chi^2_{dim(\beta)}$$

to test the null hypothesis of no difference between the two estimators.