Regression analysis with panel data: first differencing, fixed effects and random effects estimation

Bernard Lejeune

Supplemental lecture notes V

Advanced Econometrics
HEC-University of Liège
Academic year 2021-2022

- These lecture notes restate, in matrix form and with more details, the main results of Sections 13-5, 14-1 and 14-2 of Wooldridge (2016).

1. Regression and panel data

- Panel data are basically cross-sectional data where, for each individual (person, firm, city ,...) drawn at random from a given population, some dependent and explanatory variables are observed not only once (at some point in time), but repeatedly over a certain number of time periods. Typically, the number N of observed individuals is large (from tens to thousands), and the number T of observed time periods is small (from 2 to, say, 10).

- As cross-sectional data and time series data, panel data may be analyzed using a multiple linear regression model such as:

$$y_{it} = \beta_1 + \beta_2 x_{it2} + ... + \beta_k x_{itk} + u_{it}$$
$$\Leftrightarrow \quad y_{it} = X_{it}\beta + u_{it} \tag{1}$$

where $i = 1, ..., N$ indexes the individuals, $t = 1, ..., T$ is a time index, $X_{it} = (1, x_{it2}, ..., x_{itk})$ is a $1 \times k$ (row) vector of explanatory variables (including a constant, as well as usually a set of time dummies) and $\beta = (\beta_1, \beta_2, ..., \beta_k)$ is a $k \times 1$ (column) vector. Stacking (with the correct temporal ordering) the T periods of observations of an individual, for any individual i drawn from the population, we can write :

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} 1 & x_{i12} & \cdots & x_{i1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{iT2} & \cdots & x_{iTk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix}$$

$$\Leftrightarrow \quad Y_i = X_i\beta + u_i \ , \quad i = 1, ..., N$$

where Y_i and u_i are $T \times 1$ vectors, and X_i is a $T \times k$ matrix, whose the t-th row is equal to X_{it}. Stacking further the (T-variate) individual observations of a random sample of N individuals, which means a total number of observations equal to NT, we can finally write :

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} \cdots X_1 \cdots \\ \vdots \\ \cdots X_i \cdots \\ \vdots \\ \cdots X_N \cdots \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{bmatrix}$$

$$\Leftrightarrow \quad Y = X\beta + u$$

where Y and u are $NT \times 1$ vectors, and X is a $NT \times k$ matrix, whose the i-th $T \times k$ block row is equal to X_i.

- The most important feature of panel data is that they allow to evaluate partial effects – i.e., the effect of variables, the other variables being held constant – free from omitted variable bias due to unobserved time-constant individual-specific variables, so that these partial effects may more confidently be interpreted as genuine causal effects. The basic idea is to view the error u_{it} in (1) as consisting of two components : on the one hand, a component, noted a_i, standing for all time-constant individual-specific variables that affect y_{it} but have been omitted because they are unobserved, and on the other hand, a remaining error term, noted ε_{it} :

$$u_{it} = a_i + \varepsilon_{it} \tag{2}$$

The component a_i is usually called an (unobserved) individual effect or a fixed (individual) effect, and the remaining error ε_{it} is often referred to as the idiosyncratic error. With the composite error (2), (1) can be written as[1] :

$$y_{it} = X_{it}\beta + a_i + \varepsilon_{it} \tag{3}$$

where it is naturally assumed that :

$$E(\varepsilon_{it}|X_{it}, a_i) = 0 \tag{4}$$

which is equivalent to :

$$E(y_{it}|X_{it}, a_i) = X_{it}\beta + a_i \tag{5}$$

[1] Be aware : in Woodridge (2016), u_{it} is noted ν_{it} and ε_{it} is noted u_{it}, so that the composite error term is written as : $\nu_{it} = a_i + u_{it}$.

In the error component model (3), under assumption (4), or equivalently (5), the vector of parameters β measures the partial effect of the different explanatory variables on y, the other observed variables as well as any other unobserved time-constant individual-specific variables being held constant.

With cross-sectional data only, unbiased estimation of such partial effects – i.e., controlling for both observed variables and unobserved individual effects – is not possible. Being unobserved, the individual effects can not be included in the regression, so that, as a result of the usual omitted variable bias, the OLS estimator is generally biased, unless of course the individual effects are uncorrelated with the observed explanatory variables. The same omitted variable problem likewise arises with panel data if the vector of parameters β is merely estimated by pooled OLS, i.e., by merely regressing y_{it} on the observed explanatory variables $X_{it} = (1, x_{it2}, ..., x_{itk})$ using all available NT observations.

Panel data however offer ways to solve the problem. Because the individuals are observed more than once, and the individual effects are by definition time-constant, it is actually possible to get rid of the unobserved individual effects by simple transformations of the data, namely either by first differencing the data or by time-demeaning the data, so that the partial effects measured by the vector of parameters β in (5) may be estimated without requiring the individual effects to be observed. The first differencing and time-demeaning transformations are the basis of the so-called first-difference estimator and fixed effects (or within) estimator. Both estimators provide an unbiased and consistent estimator of the vector of parameters β in model (5) without any assumption – i.e., no restriction – about the way the individual effects a_i are related to the observed explanatory variables X_{it}, with however one caveat: only the parameters – and thus the partial effects – of time-varying variables can actually be estimated. This is because the effect of any time-constant variable is by construction indistinguishable from time-constant individual effects.

If the individual effects are assumed unrelated with the observed explanatory variables, as suggested above, the individual effects may be ignored and the mere pooled OLS estimator provides an unbiased and consistent estimator of the vector of parameters β in model (5). But in this case, a more efficient estimator, which takes into account the dependence in time of the errors u_{it} induced by the presence of the (assumed uncorrelated with the x_{itk}'s) individual effects a_i, may be derived. This more efficient estimator takes the form of a feasible generalized least squares (FGLS) estimator and is called the random effects estimator.

- Hereafter, we define and outline the properties of, successively, the first-difference (FD), the fixed effects (FE) and the random effects (RE) estimators of the error component model (3). Preliminarily, because it is of interest of its own, but foremost because the FD, the FE and the RE estimators actually are nothing but pooled OLS estimators applied to transformed models, we first look at the properties of pooled OLS estimation.

2. Preliminary: pooled OLS estimation

- The pooled OLS (POLS) estimator can be written as:

$$\begin{aligned}\hat{\beta}_{POLS} &= (X'X)^{-1} X'Y \\ &= \left(\sum_{i=1}^{N} X_i'X_i\right)^{-1} \sum_{i=1}^{N} X_i'Y_i \\ &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'X_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}'y_{it}\end{aligned}$$

- We first outline generic properties of the POLS estimator, ignoring the error component structure (2). We then consider pooled OLS estimation of the so-called 'random effects model', i.e., the error component model (3) where it is assumed that the (unobserved) individual effects a_i are unrelated with the observed explanatory variables X_{it}.

2.1. Generic properties of the pooled OLS estimator

- Hereafter, we show that the pooled OLS estimator $\hat{\beta}_{POLS}$ has the same finite sample and asymptotic properties as the OLS estimator in the cross-sectional case under the following generic assumptions:

  – POLS.1 Linearity in parameters

    The model can be written as:

$$y_{it} = X_{it}\beta + u_{it} \ , \quad i = 1, ..., N, \ t = 1, ..., T$$
$$\Leftrightarrow \quad Y_i = X_i\beta + u_i \ , \quad i = 1, ..., N$$

    where β is a $k \times 1$ vector of unknown parameters and u_i is a $T \times 1$ error term.

  – POLS.2 Random sampling

    The available data are realizations of a random sample of size N, $\{(X_i, Y_i): i = 1, ..., N\}$, following the model in assumption POLS.1.

  – POLS.3 No perfect collinearity

    In the sample (and thus in the population), there is no exact linear relationship among the explanatory variables (including the constant).

  – POLS.4 Zero conditional mean

    For each i, the expected value of u_i given any values of X_i is equal to zero, which is equivalent to say that the expected value of Y_i given any values of X_i is equal to $X_i\beta$:

$$E(u_i|X_i) = 0 \ \Leftrightarrow \ E(Y_i|X_i) = X_i\beta \ , \quad i = 1, ..., N$$

  – POLS.5 Homoskedasticity and no serial correlation

For each i, the matrix of variance-covariance of u_i given any values of X_i is constant and equal to $\sigma^2 I_T$, which is equivalent to say that the matrix of variance-covariance of Y_i given any values of X_i is constant and equal to $\sigma^2 I_T$:

$$V(u_i|X_i) = \sigma^2 I_T \;\Leftrightarrow\; V(Y_i|X_i) = \sigma^2 I_T \;,\;\; i = 1, ..., N$$

where σ^2 is an unknown parameter[2].

– POLS.6 Normality

For each i, the distribution of u_i is the same given any values of X_i – i.e., u_i is independent of X_i – and is jointly normal with zero mean and variance-covariance equal to $\sigma^2 I_T$, which is equivalent to say that the distribution of Y_i given any value of X_i is jointly normal with mean equal to $X_i\beta$ and variance-covariance equal to $\sigma^2 I_T$:

$$u_i|X_i \sim N(0, \sigma^2 I_T) \;\Leftrightarrow\; Y_i|X_i \sim N(X_i\beta, \sigma^2 I_T) \;,\;\; i = 1, ..., N$$

- In a nutshell, assumptions POLS.1 – POLS.6 are a multivariate (T-variate) version of the cross-sectional assumptions MLR.1 – MLR.6. As a matter of fact, the cross-sectional assumptions MLR.1 – MLR.6 are just a special case of assumptions POLS.1 – POLS.6, where $T = 1$. This multivariate version deserves a few comments :

  – In detailed form, for each i, assumption POLS.4 must be read : $E(y_{it}|X_i) = E(y_{it}|X_{it}) = X_{it}\beta \Leftrightarrow E(u_{it}|X_i) = E(u_{it}|X_{it}) = 0$, for all $t = 1, ..., T$. In other words, it is assumed that the explanatory variables are strictly exogenous.

  – By the law of iterated expectations, assumption POLS.4 implies, for each i, that the unconditional mean of u_{it} is zero (i.e., $E(u_{it}) = 0$, for all $t = 1, ..., T$), and that u_{it} is uncorrelated (have zero covariance) with each explanatory variable, in each time period (i.e., $E(x_{isj}u_{it}) = 0$, for all $s, t = 1, ..., T$, $j = 2, ..., k$).

  – In detailed form, for each i, assumption POLS.5 must be read : $Var(u_{it}|X_i) = \sigma^2 \Leftrightarrow Var(y_{it}|X_i) = \sigma^2$, for all $t = 1, ..., T$, and $Cov(u_{it}, u_{is}|X_i) = 0 \Leftrightarrow Cov(y_{it}, y_{is}|X_i) = 0$, for all $t \neq s$. It thus assumes both constant variance and no correlation across time of the errors, conditional on the explanatory variables (in all time periods) X_i.

  – As in the cross-sectional case, assumptions POLS.5 and POLS.6 are auxiliary assumptions. They are only crucial for exact in finite sample inference.

2.1.1. Finite sample properties of POLS

- As in the cross-sectional case, because the random sampling assumption POLS.2 implies independence of the observations across i, assumptions POLS.1 – POLS.6

[2] I_T denotes a $T \times T$ identity matrix.

imply the so-called 'classical linear model' assumptions E.1 – E.5 stated in the supplemental lecture notes I (hereafter SLN-I)[3]. Under assumptions POLS.1 – POLS.6, the POLS estimator has thus the same finite sample properties as the OLS estimator in the cross-sectional case. More specifically :

- Under assumptions POLS.1 – POLS.4 (linearity in parameters, random sampling, no perfect colinearity and zero conditional mean), which imply assumptions E.1 – E.3, from Property 1 in SLN-I, $\hat{\beta}_{POLS}$ is an unbiased estimator of β.
- If assumption POLS.5 (homoskedasticity and no serial correlation) is added to assumptions POLS.1 – POLS.4, so that assumptions E.1 – E.4 hold, from Property 2 and 3 in SLN-I, the (conditional) variance-covariance matrix of $\hat{\beta}_{POLS}$ is given by $V(\hat{\beta}_{POLS}|X) = \sigma^2 (X'X)^{-1}$ and $\hat{\beta}_{POLS}$ is the best linear unbiased estimator (BLUE) of β. Also, from Property 5 in SLN-I, $\hat{s}^2 = \frac{1}{NT-k} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2 = \frac{\hat{u}'\hat{u}}{NT-k}$ is an unbiased estimator of σ^2.
- If assumption POLS.6 (normality) is added to assumptions POLS.1 – POLS.5, so that assumptions E.1 – E.5 hold, from Property 4 in SLN-I, the (conditional) distribution of $\hat{\beta}$ is normal and given by $\hat{\beta}_{POLS}|X \sim N(\beta, \sigma^2(X'X)^{-1})$.

- Needless to say, as under assumptions POLS.1 – POLS.6 the finite sample properties of the POLS estimator $\hat{\beta}_{POLS}$ are the same as the finite sample properties of the OLS estimator in the cross-sectional case, all usual inference procedures derived in Section 4.1, 4.2 and 4.4.1 of SLN-I[4] – i.e., the confidence intervals for β_j or a single linear combination $R_0\beta$, the two-sided and one-sided t-tests of β_j or a single linear combination $R_0\beta$, and the F-test of multiple linear restrictions – are of course valid and exact in finite sample.

2.1.2. Asymptotic properties of POLS

- In the present panel data context, asymptotic properties refer to sampling distribution properties which are valid when the number of sampled individuals N goes to infinity, while the number T of time periods over which the individuals are observed remains fixed (i.e., $N \to \infty, T$ fixed).

A. Consistency of $\hat{\beta}_{POLS}$

- We have the following property :

Property 16 Consistency of $\hat{\beta}_{POLS}$

Under assumptions POLS.1 – POLS.4, the POLS estimator $\hat{\beta}_{POLS}$ is a consistent estimator of β :

$$\hat{\beta}_{POLS} \xrightarrow{p} \beta$$

---

[3] i.e., assumptions E.1 – E.5 hold whenever assumptions POLS.1 – POLS.6 hold. Note that the converse is not true.

[4] The only notable difference is that the number of observations n is now equal to NT, so that the number of degrees of freedom $n - k$ is now equal to $NT - k$.

A sketch of the proof is the same as for the consistency of the OLS estimator in the cross-sectional case (Property 6 in SLN-I) : the only notable difference is that here X_i refers to a $T \times k$ matrix (rather than a $1 \times k$ (row) vector), and Y_i and u_i to $T \times 1$ vectors (rather than scalars). Hereafter is a concise version of it. Under assumptions POLS.1 – POLS.3, we have :

$$\hat{\beta}_{POLS} = \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' Y_i\right) = \beta + \left(\frac{1}{N}\sum_{i=1}^{N} X_i' X_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} X_i' u_i\right) \quad (6)$$

Under random sampling, both $X_i'X_i$ and $X_i'u_i$ are i.i.d. across i, so that $\frac{1}{N}\sum_{i=1}^{N} X_i'X_i$ and $\frac{1}{N}\sum_{i=1}^{N} X_i'u_i$ are both sample average to which the law of large numbers[5] (LLN) can be applied. From the LLN, we have :

$$\frac{1}{N}\sum_{i=1}^{N} X_i' X_i \xrightarrow{p} E(X_i'X_i) = A$$

Under the zero conditional mean assumption POLS.4 $E(u_i|X_i) = 0$, by the law of iterated expectations, we have :

$$E(X_i'u_i) = E\left[E(X_i'u_i|X_i)\right] = E\left[X_i' E(u_i|X_i)\right] = E\left[X_i' \cdot 0\right] = 0 \quad (7)$$

From the LLN, we thus also have :

$$\frac{1}{N}\sum_{i=1}^{N} X_i' u_i \xrightarrow{p} 0$$

so that, from (6), we finally have :

$$\hat{\beta}_{POLS} \xrightarrow{p} \beta + A^{-1} \cdot 0 = \beta$$

B. Asymptotic normality of $\hat{\beta}_{POLS}$ and inference

- We have the following property :

Property 17 Asymptotic normality of $\hat{\beta}_{POLS}$
Under assumptions POLS.1 – POLS.5, the POLS estimator $\hat{\beta}_{POLS}$ is asymptotically normally distributed :

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \xrightarrow{d} N(0, \sigma^2 A^{-1}), \quad \text{where } A = E(X_i'X_i) \quad (8)$$

A sketch of the proof is as follows. It is again basically the same as for the OLS estimator in the cross-sectional case (Property 7 in SLN-I). Hereafter is a concise version of it. Under assumptions POLS.1 – POLS.3, from (6), we have :

$$\hat{\beta}_{POLS} = \beta + \left(\frac{1}{N}\sum_{i=1}^{N} X_i' X_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} X_i' u_i\right)$$

[5] If $\{Z_i \colon i = 1, ..., n\}$ are i.i.d. random variables with $E(Z_i) = m$, then by the LLN we have : $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} m$.

$$\Leftrightarrow \sqrt{N}(\hat{\beta}_{POLS} - \beta) = \left(\frac{1}{N}\sum_{i=1}^{N} X_i' X_i\right)^{-1} \left(N^{-\frac{1}{2}}\sum_{i=1}^{N} X_i' u_i\right)$$

Under random sampling, both $X_i'X_i$ and $X_i'u_i$ are i.i.d. across i. As already outlined, from the LLN, we have $\frac{1}{N}\sum_{i=1}^{N} X_i'X_i \xrightarrow{p} E(X_i'X_i) = A$, and we can write:

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \overset{as}{=} A^{-1} \left(N^{-\frac{1}{2}}\sum_{i=1}^{N} X_i' u_i\right) \tag{9}$$

so that $\sqrt{N}(\hat{\beta}_{POLS} - \beta)$ has asymptotically the same distribution as $A^{-1}\left(N^{-\frac{1}{2}}\sum_{i=1}^{N} X_i'u_i\right)$. As also already outlined, under the zero conditional mean assumption POLS.4, we have $E\left(X_i'u_i\right) = 0$. From the central limit theorem[6] (CLT), we thus have:

$$N^{-\frac{1}{2}}\sum_{i=1}^{N} X_i' u_i \xrightarrow{d} N(0, B), \quad \text{where } B = V(X_i'u_i) = E(X_i'u_iu_i'X_i)$$

so that:

$$A^{-1}\left(N^{-\frac{1}{2}}\sum_{i=1}^{N} X_i' u_i\right) \xrightarrow{d} N(0, A^{-1}BA^{-1})$$

and thus, from (9):

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \xrightarrow{d} N(0, A^{-1}BA^{-1}) \tag{10}$$

To complete the proof, it remains to show that, under the homoskedasticity and no serial correlation assumption POLS.5, we have $B = \sigma^2 A$. Under assumption POLS.5, by the law of iterated expectations, we have:

$$\begin{aligned} B &= V(X_i'u_i) = E(X_i'u_iu_i'X_i) \\ &= E\left[E(X_i'u_iu_i'X_i|X_i)\right] = E\left[X_i'E(u_iu_i'|X_i)X_i\right] \\ &= E\left[X_i'(\sigma^2 I_T)X_i\right] = \sigma^2 E(X_i'X_i) = \sigma^2 A \end{aligned}$$

so that, from (10), we finally have:

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \xrightarrow{d} N(0, \sigma^2 A^{-1})$$

- The limiting distributional result (8) is the same as the limiting distributional result given by Property 7 in SLN-I for the OLS estimator in the cross-sectional case[7]. As in the cross-sectional case, it provides an approximate finite sample distribution for the POLS estimator $\hat{\beta}$:

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \approx N(0, \sigma^2 A^{-1})$$

$$\Leftrightarrow \quad \hat{\beta}_{POLS} \approx N(\beta, \sigma^2 A^{-1}/N) \tag{11}$$

[6] if $\{Z_i\colon\, i = 1, ..., n\}$ are i.i.d. $(k \times 1)$ random vectors with $E(Z_i) = m$ and $V(Z_i) = \Sigma$, then by the CLT we have: $\sqrt{n}(\bar{Z}_n - m) = n^{-\frac{1}{2}}\sum_{i=1}^{n}(Z_i - m) \xrightarrow{d} N(0, \Sigma)$.

[7] As a matter of fact, Property 7 in SLN-I is just a special case of Property 17, where $T = 1$.

which can be used – when N is sufficiently large – for performing inference (confidence interval, hypothesis testing) without having to rely on any other assumption than assumptions POLS.1 – POLS.5, i.e., in particular without having to rely on any normality assumption.

- As in the cross-section case, an estimator of the asymptotic variance $Avar(\hat{\beta}_{POLS}) = \sigma^2 A^{-1}/N$ is simply obtained by replacing σ^2 and A by consistent estimators. From the LLN, a consistent estimator of $A = E(X_i'X_i)$ is given by $\frac{1}{N}\sum_{i=1}^N X_i'X_i = X'X/N$, and a consistent estimator of σ^2 is likewise still given by $\hat{s}^2 = \frac{1}{NT-k}\sum_{i=1}^N\sum_{t=1}^T \hat{u}_{it}^2$, where $\hat{u}_{it} = y_{it} - X_{it}\hat{\beta}_{POLS}$, so that an estimator of $Avar(\hat{\beta}_{POLS}) = \sigma^2 A^{-1}/N$ is given by:

$$\hat{V}(\hat{\beta}_{POLS}) = \hat{s}^2\,(X'X)^{-1} \tag{12}$$

- As the limiting distributional result (8), or equivalently the approximate distributional result (11), and the estimator (12) of the asymptotic variance $Avar(\hat{\beta}_{POLS})$ are the same as in the cross-sectional case, from the results derived in Section 4.3 and Section 4.4.2 of SLN-I, it follows that all usual inference procedures[8] – confidence interval for β_j or a single linear combination $R_0\beta$, two-sided and one-sided t-tests of β_j or a single linear combination $R_0\beta$, F-test (or Wald test) of multiple linear restrictions – are likewise asymptotically valid – i.e., approximately valid for N sufficiently large – in the present panel data context under assumptions POLS.1 – POLS.5.

C. Heteroskedasticity and autocorrelation robust inference

- The outlined above asymptotic properties of the POLS estimator do not rely on any normality assumption, but they require homoskedasticity and no serial correlation. If the homoskedasticity and no serial correlation assumption POLS.5 does not hold, then the POLS estimator is still unbiased and consistent, but all usual inference procedures are no longer valid. However, it is again possible to derive inference procedures which are valid in the presence of both heteroskedasticity and serial correlation of unknown form, i.e., which are robust to arbitrary form of both heteroskedasticity and autocorrelation.

- Heteroskedasticity and autocorrelation robust inference procedures basically rely on the following property, which outlines the asymptotic properties of the POLS estimator $\hat{\beta}_{POLS}$ under only assumptions POLS.1 – POLS.4 (Linearity in parameters, random sampling, no perfect collinearity and zero conditional mean):

Property 18 Asymptotic properties of $\hat{\beta}_{POLS}$ without homoskedasticity and no serial correlation
Under assumptions POLS.1 – POLS.4, the POLS estimator $\hat{\beta}_{POLS}$ is a consistent estimator of β:

$$\hat{\beta}_{POLS} \xrightarrow{p} \beta$$

and is asymptotically normally distributed as:

[8] Note that the appropriate number of degrees of freedom is here $NT - k$ (and not $N - k$).

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \xrightarrow{d} N(0, A^{-1}BA^{-1}) \quad (13)$$

where:

$$A = E(X_i'X_i) \quad \text{and} \quad B = E(X_i'u_iu_i'X_i)$$

The fact that $\hat{\beta}_{POLS}$ is consistent for β under assumptions POLS.1 – POLS.4 was already outlined in Property 16. On the other hand, the limiting distribution result (13) has already be shown to hold likewise under assumptions POLS.1 – POLS.4 in the sketch of the proof of Property 17: see the intermediary result (10).

- The limiting distribution result (13) is a direct extension of the limiting distributional result given by Property 9 in the supplemental lecture notes III (hereafter SLN-III) for the OLS estimator in the cross-sectional case[9]. As in the cross-sectional case, it provides an approximate finite sample distribution for the POLS estimator $\hat{\beta}$:

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \xrightarrow{d} N(0, A^{-1}BA^{-1})$$

$$\Leftrightarrow \quad \hat{\beta}_{POLS} \approx N(\beta, A^{-1}BA^{-1}/N) \quad (14)$$

which can be used – when N is sufficiently large – for performing robust inference (confidence interval, hypothesis testing) without having to rely on any other assumption than assumptions POLS.1 – POLS.4, i.e., in particular without having to rely on the homoskedasticity and no serial correlation assumption POLS.5.

- For inference based on the limiting distributional result (13), or equivalently on the approximate distributional result (14), we need an estimator of the asymptotic variance $Avar(\hat{\beta}_{POLS}) = A^{-1}BA^{-1}/N$. This requires consistent estimators of A and B. We already know that $X'X/N$ is a consistent estimator of A (this directly follows from the LLN). A consistent estimator of $B = E(X_i'u_iu_i'X_i)$ is simply given by $\frac{1}{N}\sum_{i=1}^{N} X_i'\hat{u}_i\hat{u}_i'X_i$, where obviously $\hat{u}_i = Y_i - X_i\hat{\beta}_{POLS}$. This is formalized in the following property[10]:

Property 19 Consistent estimator of B
Under POLS.1 – POLS.4, $\frac{1}{N}\sum_{i=1}^{N} X_i'\hat{u}_i\hat{u}_i'X_i$ is a consistent estimator of B:

$$\frac{1}{N}\sum_{i=1}^{N} X_i'\hat{u}_i\hat{u}_i'X_i \xrightarrow{p} B$$

Here is the intuition of this property: from the LLN, $\frac{1}{N}\sum_{i=1}^{N} X_i'u_iu_i'X_i \xrightarrow{p} E(X_i'u_iu_i'X_i) = B$. As $\hat{u}_i$ converges to u_i (because $\hat{\beta}_{POLS} \xrightarrow{p} \beta$), we also have

[9] As a matter of fact, Property 9 in SLN-III is just a special case of Property 18, where $T = 1$.
[10] For more details, see Wooldridge (2010) p. 171-172. Note that Wooldridge (2010) relies on weaker assumptions than those considered here.

$\frac{1}{N}\sum_{i=1}^{N} X_i'\hat{u}_i\hat{u}_i'X_i \xrightarrow{p} B.$

- Using the outlined above consistent estimators of A and B, an estimator of $Avar(\hat{\beta}_{POLS}) = A^{-1}BA^{-1}/N$ is given by:

$$\hat{V}_R(\hat{\beta}_{POLS}) = (X'X)^{-1}\left(\sum_{i=1}^{N} X_i'\hat{u}_i\hat{u}_i'X_i\right)(X'X)^{-1} \tag{15}$$

where the subscript 'R' simply stands for 'Robust'[11]. As usual, the diagonal elements $V\hat{a}r_R(\hat{\beta}_{POLS_j})$ of the $k \times k$ matrix estimator $\hat{V}_R(\hat{\beta}_{POLS})$ being the estimators of the variance $Avar(\hat{\beta}_{POLS_j})$ of the estimator $\hat{\beta}_{POLS_j}$ of the different parameters β_j $(j = 1, ..., k)$, natural estimators of the asymptotic standard error $As.e.(\hat{\beta}_{POLS_j}) = \sqrt{Avar(\hat{\beta}_{POLS_j})}$ of the estimator $\hat{\beta}_{POLS_j}$ of the different parameters β_j, as well as a natural estimator of the asymptotic standard error $As.e.(R_0\hat{\beta}_{POLS}) = \sqrt{Avar(R_0\hat{\beta}_{POLS})} = \sqrt{R_0 Avar(\hat{\beta}_{POLS})R_0'}$ of the estimator $R_0\hat{\beta}_{POLS}$ of a single linear combination $R_0\beta$ of β, are likewise given by:

$$s.\hat{e}._R(\hat{\beta}_{POLS_j}) = \sqrt{V\hat{a}r_R(\hat{\beta}_{POLS_j})}, \quad j = 1, ..., k \tag{16}$$

and:

$$s.\hat{e}._R(R_0\hat{\beta}_{POLS}) = \sqrt{R_0\hat{V}_R(\hat{\beta}_{POLS})R_0'} \tag{17}$$

where R_0 is a $1 \times k$ (row) vector of constants.

- As for heteroskedasticity robust inference in the cross-sectional case, the limiting distributional result (13), or equivalently the approximate distributional result (14), and the robust estimators $\hat{V}_R(\hat{\beta}_{POLS})$, $s.\hat{e}._R(\hat{\beta}_{POLS_j})$ and $s.\hat{e}._R(R_0\hat{\beta}_{POLS})$ given above in respectively (15), (16) and (17), provide all which is needed for performing heteroskedasticity and autocorrelation robust inference after POLS estimation. Following exactly the same reasoning as in Section 4.3 and Section 4.4.2 of SLN-I, it may readily be checked that if in all the usual inference procedures[12] – confidence interval for β_j or a single linear combination $R_0\beta$, two-sided and one-sided t-tests of β_j or a single linear combination $R_0\beta$, F-test (or Wald test) of multiple linear restrictions – we replace the usual estimators $\hat{V}(\hat{\beta}_{POLS})$, $s.\hat{e}.(\hat{\beta}_{POLS_j})$ and $s.\hat{e}.(R_0\hat{\beta}_{POLS})$ by their robust versions $\hat{V}_R(\hat{\beta}_{POLS})$, $s.\hat{e}._R(\hat{\beta}_{POLS_j})$ and $s.\hat{e}._R(R_0\hat{\beta}_{POLS})$, then we obtain inference procedures that are asymptotically valid – i.e., approximately valid for N sufficiently large – under only assumptions POLS.1 – POLS.4, i.e., without having to rely on the homoskedasticity and no serial correlation assumption POLS.5 (as well as on any normality assumption).

[11] This robust estimator, which is a direct extension of the heretoskedasticity robust estimator outlined in SLN-III for the cross-sectional case, is sometimes referred to as 'panel robust' or 'cluster robust'. Most often, it is simply referred to as 'robust'.

[12] Note again that the appropriate number of degrees of freedom is here $NT - k$ (and not $N - k$).

- Remarks :

  – Modern econometric software provide special options to compute robust standard errors and perform robust tests after pooled OLS estimation.

  – The asymptotic results outlined above (i.e., Properties 16 – 19) actually hold under weaker assumptions than those considered here[13]. In particular, it may be shown that Properties 16, 18 and 19 still hold if the zero conditional mean assumption POLS.4 is replaced by the weaker cross-sectional like assumption[14] :

  POLS.4' : for each i and t, $E(u_{it}|X_{it}) = 0 \Leftrightarrow E(y_{it}|X_{it}) = X_{it}\beta$

  i.e., Properties 16, 18 and 19 still hold without actually having to assume that the explanatory variables are strictly exogenous. In a nutshell, under assumption POLS.4', by the law of iterated expectations, for each i and t, we have :

$$\begin{aligned} E(X'_{it}u_{it}) &= E\left[E(X'_{it}u_{it}|X_{it})\right] \\ &= E\left[X'_{it}E(u_{it}|X_{it})\right] = E\left[X'_{it} \cdot 0\right] = 0 \end{aligned}$$

  so that :

$$E\left(X'_i u_i\right) = E\left(\sum_{t=1}^{T} X'_{it}u_{it}\right) = \sum_{t=1}^{T} E(X'_{it}u_{it}) = 0, \quad i = 1, ..., N$$

  which actually is the key condition – implied by assumption POLS.4 – for Properties 16, 18 and 19 to hold. Pooled OLS thus provides consistent estimation and robust inference without actually requiring strict exogeneity of the explanatory variables.

  – Be warned : the strict exogeneity assumption can not be relaxed without losing the outlined finite sample properties of the POLS estimator. For these finite sample properties to hold, assumptions POLS.1 – POLS.6 are required.

  – Note finally that none of the outlined properties of the POLS estimator requires restrictions on the dependence of the data – such as weak dependence – in the time dimension. This is basically because they rely on LLN and CLT for independent data with $N \to \infty$ and T fixed[15].

2.1.3. POLS estimation of the random effect model

- We now consider pooled OLS estimation of the so-called 'random effects model', i.e., in a nutshell, the error component model (3) where it is assumed that the (unobserved) individual effects a_i are unrelated with the observed explanatory

[13] See Wooldridge (2010), Section 7.8 for a detailed discussion.

[14] Likewise, it may be shown that Property 17 also still holds if, besides replacing assumption POLS.4 by POLS.4', assumption POLS.5 is similarly replaced by the weaker time-series like assumption POLS.5' : for each i, $Var(u_{it}|X_{it}) = \sigma^2 \Leftrightarrow Var(y_{it}|X_{it}) = \sigma^2$, for all $t = 1, ..., T$, and $E(u_{it}u_{is}|X_{it}, X_{is}) = 0$, for all $t \neq s$.

[15] For more details, see Wooldridge (2010), Section 7.8.

variables X_{it}. Formally, the random effects model assumes that the following assumptions hold[16] :

– EC.1 The model can be written as:

$$y_{it} = X_{it}\beta + a_i + \varepsilon_{it} \ , \ \ i = 1, ..., N, \ t = 1, ..., T$$
$$\Leftrightarrow \quad Y_i = X_i\beta + a_i i_T + \varepsilon_i \ , \ \ i = 1, ..., N$$

where β is a $k \times 1$ vector of unknown parameters, i_T is a $T \times 1$ vector of ones, a_i is an (unobserved) individual effect and ε_i is a $T \times 1$ error term.

– EC.2 The available data are realizations of a random sample of size N, $\{(X_i, Y_i)$: $i = 1, ..., N\}$, following the model in assumption EC.1.

– EC.3 In the sample (and thus in the population), there is no exact linear relationship among the explanatory variables (including the constant).

– EC.4 (a) $E(\varepsilon_i|X_i, a_i) = 0 \Leftrightarrow E(Y_i|X_i, a_i) = X_i\beta + a_i i_T \ , \ \ i = 1, ..., N$

(b) $E(a_i|X_i) = E(a_i) = 0 \ , \ \ i = 1, ..., N$

– EC.5 (a) $V(\varepsilon_i|X_i, a_i) = \sigma_\varepsilon^2 I_T \Leftrightarrow V(Y_i|X_i, a_i) = \sigma_\varepsilon^2 I_T \ , \ \ i = 1, ..., N$

(b) $Var(a_i|X_i) = \sigma_a^2 \ , \ \ i = 1, ..., N$

where σ_ε^2 and σ_a^2 are unknown parameters.

- The above set of assumptions deserves the following comments:

  – Assumptions EC.1 – EC.3 are the same as assumptions POLS.1 – POLS.3, where it is assumed that the error u_{it} consists of two components: an (unobserved) individual effect a_i and a remaining error term ε_{it}, i.e., $u_{it} = a_i + \varepsilon_{it}$ (in vector form, $u_i = a_i i_T + \varepsilon_i$).

  – In detailed form, for each i, assumption EC.4a must be read: $E(y_{it}|X_i, a_i) = E(y_{it}|X_{it}, a_i) = X_{it}\beta + a_i \Leftrightarrow E(\varepsilon_{it}|X_i, a_i) = E(\varepsilon_{it}|X_{it}, a_i) = 0$, for all $t = 1, ..., T$. In other words, it is assumed that the explanatory variables are strictly exogenous conditional on the individual effects. As discussed in the supplemental lecture notes IV (hereafter SLN-IV), this implies that neither feedback from the current value of the dependent variable to the future values of the explanatory variables nor lagged dependent variables are allowed.

  – Assumption EC.4b is the distinctive assumption of the random effects model: it states that the individual effects a_i are unrelated with – statistically, mean independent of – the observed explanatory variables (in all time periods) X_i. The assumption that $E(a_i) = 0$ is without loss of generality, provided that an intercept is included in X_{it}.

[16] The abbreviation 'EC' stands for 'Error Component'. The assumptions listed here and in the following sections are the same as those listed in Appendix 13A and Appendix 14A of Wooldridge (2016), but are organized in a different (simpler and more concise) way.

– By the law of iterated expectations, assumption EC.4a implies, for each i, that the unconditional mean of ε_{it} is zero (i.e., $E(\varepsilon_{it}) = 0$, for all $t = 1, ..., T$), that ε_{it} is uncorrelated (have zero covariance) with each explanatory variable, in each time period (i.e., $E(x_{isj}\varepsilon_{it}) = 0$, for all $s, t = 1, ..., T$, $j = 2, ..., k$), and that ε_{it} is uncorrelated with a_i (i.e., $E(a_i\varepsilon_{it}) = 0$, for all $t = 1, ..., T$). On the other hand, assumption EC.4b implies, for each i, that a_i is uncorrelated (have zero covariance) with each explanatory variable, in each time period (i.e., $E(x_{itj}a_i) = 0$, for all $t = 1, ..., T$, $j = 2, ..., k$).

– In detailed form, for each i, assumption EC.5a must be read : $Var(\varepsilon_{it}|X_i, a_i) = \sigma_\varepsilon^2 \Leftrightarrow Var(y_{it}|X_i, a_i) = \sigma_\varepsilon^2$, for all $t = 1, ..., T$, and $Cov(\varepsilon_{it}, \varepsilon_{is}|X_i, a_i) = 0 \Leftrightarrow Cov(y_{it}, y_{is}|X_i, a_i) = 0$, for all $t \neq s$. It thus assumes both constant variance and no correlation across time of the errors ε_{it}, conditional on the explanatory variables (in all time periods) X_i and the individual effects a_i. On the other hand, assumption EC.5b assumes constant variance of the individual effects a_i, conditional on the explanatory variables (in all time periods) X_i.

- Under assumption EC.4a, by the law of iterated expectations, we have :

$$E(\varepsilon_i|X_i) = E\left[E(\varepsilon_i|X_i, a_i)|X_i\right] = E\left[0\,|X_i\right] = 0 \tag{18}$$

and :

$$\begin{aligned} E(a_i\varepsilon_i|X_i) &= E\left[E(a_i\varepsilon_i|X_i, a_i)|X_i\right] \\ &= E\left[a_iE(\varepsilon_i|X_i, a_i)|X_i\right] = E\left[a_i \cdot 0|X_i\right] = 0 \end{aligned}$$

so that, if assumption EC.4b also holds, for the composite error term $u_i = a_ii_T + \varepsilon_i$, we have :

$$E(u_i|X_i) = i_TE(a_i|X_i) + E(\varepsilon_i|X_i) = i_T \cdot 0 + 0 = 0$$

If in addition assumption EC.5a holds, by the law of iterated expectations, we have:

$$\begin{aligned} V(\varepsilon_i|X_i) &= E(\varepsilon_i\varepsilon_i'|X_i) = E\left[E(\varepsilon_i\varepsilon_i'|X_i, a_i)|X_i\right] \\ &= E\left[V(\varepsilon_i|X_i, a_i)|X_i\right] = E\left[\sigma_\varepsilon^2 I_T|X_i\right] = \sigma_\varepsilon^2 I_T \end{aligned} \tag{19}$$

so that, if assumption EC.5b also holds, for the composite error term $u_i = a_ii_T + \varepsilon_i$, we also have :

$$\begin{aligned} V(u_i|X_i) &= E(u_iu_i'|X_i) = E[(a_ii_T + \varepsilon_i)(a_ii_T + \varepsilon_i)'|X_i)] \\ &= i_Ti_T'E(a_i^2|X_i) + i_TE(a_i\varepsilon_i'|X_i) + E(a_i\varepsilon_i|X_i)i_T' + E(\varepsilon_i\varepsilon_i'|X_i) \\ &= i_Ti_T'Var(a_i|X_i) + 0 + 0 + V(\varepsilon_i|X_i) \\ &= \sigma_a^2 J_T + \sigma_\varepsilon^2 I_T \end{aligned}$$

where J_T stands for a $T \times T$ matrix of ones, i.e., in detailed form :

$$V(u_i|X_i) = \begin{bmatrix} \sigma_a^2 + \sigma_\varepsilon^2 & \sigma_a^2 & \cdots & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_\varepsilon^2 & \cdots & \sigma_a^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_a^2 & \sigma_a^2 & \cdots & \sigma_a^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

- According to the above discussion, under the random effects model assumptions EC.1 – EC.5, the generic POLS assumptions POLS.1 – POLS.4 (linearity in parameters, random sampling, no perfect colinearity and zero conditional mean) hold, but, by definition, the POLS assumption POLS.5 (homoskedasticity and no serial correlation) does not hold. As a consequence, from the generic properties of the POLS estimator:

  - Under assumptions EC.1 – EC.4, which imply assumptions POLS.1 – POLS.4, the POLS estimator $\hat{\beta}_{POLS}$ is an unbiased estimator of β. However, because assumption POLS.5 does not hold, neither $\hat{\beta}_{POLS}$ is the best linear unbiased estimator (BLUE) of β, nor the usual inference procedures are valid (even asymptotically).

  - Under the same assumptions EC.1 – EC.4, the POLS estimator $\hat{\beta}_{POLS}$ is also consistent for β, and asymptotically valid – i.e., approximately valid for N sufficiently large – inference procedures are provided by the heteroskedasticity and autocorrelation robust inference procedures outlined in Section 2.1.2 - C.

- Remarks:

  - The best linear unbiased estimator of β under the random effects model assumptions EC.1 – EC.5, which properly accounts for the dependence in time of the errors u_{it} induced by the presence of the individual effects, is given by the so-called random effects estimator which will be considered later in Section 5.

  - The consistency – but not the unbiasedness – of $\hat{\beta}_{POLS}$ and the asymptotic validity of the robust inference procedures outlined above actually hold under weaker assumptions than those considered here[17]. In particular, according to the second remark made at the end of Section 2.1.2 - C, it may be shown that they still hold if assumption EC.4 is replaced by the weaker cross-sectional like assumption:

    EC.4': for each i and t,

    (a) $E(\varepsilon_{it}|X_{it}, a_i) = 0 \Leftrightarrow E(Y_{it}|X_{it}, a_i) = X_{it}\beta + a_i$

    (b) $E(a_i|X_{it}) = E(a_i) = 0$

    i.e., they still hold without actually having to assume that the explanatory variables are strictly exogenous conditional on the individual effects, with however one caveat: if feedback from the current value of the dependent

[17] See Wooldridge (2010), Section 10.3 for a detailed discussion.

variable to the future values of the explanatory variables is allowed under assumption EC.4', lagged dependent variables are still not allowed (because with X_{it} containing a lagged dependent variable, assumption EC.4' (b) can not be true[18]).

3. First differencing estimation

- When, in the error component model described in assumption EC.1:

$$y_{it} = X_{it}\beta + a_i + \varepsilon_{it} \ , \ \ i = 1, ..., N, \ t = 1, ..., T$$
$$\Leftrightarrow \quad Y_i = X_i\beta + a_i i_T + \varepsilon_i \ , \ \ i = 1, ..., N \tag{20}$$

both the random sampling assumption EC.2 and the zero conditional mean assumption EC.4a:

$$E(\varepsilon_i|X_i, a_i) = 0 \ \Leftrightarrow \ E(Y_i|X_i, a_i) = X_i\beta + a_i i_T \ , \ i = 1, ..., N$$

are maintained, but nothing – i.e., no restriction – is assumed regarding the relationship between the (unobserved) individual effects a_i and the observed explanatory variables X_{it} – i.e., assumption EC.4b $E(a_i|X_i) = 0$ is not maintained –, then the model is usually referred to as a 'fixed effects model' (or, less commonly, an 'unobserved effects model').

- When assumption EC.4b does not hold, the pooled OLS estimator of model (20) is generally biased, due to the usual omitted variable bias. This problem can however be circumvented by getting rid of the problematic individual effects a_i through a very simple transformation of the data, namely by first differencing the data. Let D stands for the $(T-1) \times T$ first differencing transformation matrix:

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

Because $Di_T = 0$, premultiplying both sides of (20) by D yields the transformed model:

$$\Delta Y_i = \Delta X_i\beta + \Delta\varepsilon_i \ , \ \ i = 1, ..., N \tag{21}$$

where $\Delta Y_i = DY_i$ and $\Delta\varepsilon_i = D\varepsilon_i$ are $(T-1) \times 1$ vectors equal to:

$$\Delta Y_i = \begin{bmatrix} \Delta y_{i2} \\ \vdots \\ \Delta y_{iT} \end{bmatrix} = \begin{bmatrix} y_{i2} - y_{i1} \\ \vdots \\ y_{iT} - y_{iT-1} \end{bmatrix} \text{ and } \Delta\varepsilon_i = \begin{bmatrix} \Delta\varepsilon_{i2} \\ \vdots \\ \Delta\varepsilon_{iT} \end{bmatrix} = \begin{bmatrix} \varepsilon_{i2} - \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} - \varepsilon_{iT-1} \end{bmatrix}$$

[18] See Wooldridge (2010), Section 10.3.

and $\Delta X_i = DX_i$ is a $(T-1) \times k$ matrix equal to:

$$\Delta X_i = \begin{bmatrix} \Delta X_{i2} \\ \vdots \\ \Delta X_{iT} \end{bmatrix} = \begin{bmatrix} X_{i2} - X_{i1} \\ \vdots \\ X_{iT} - X_{iT-1} \end{bmatrix}$$

i.e., in detailed form:

$$\Delta y_{it} = \Delta X_{it}\beta + \Delta\varepsilon_{it} \ , \quad i = 1, ..., N, \ t = 2, ..., T$$

where $\Delta y_{it} = y_{it} - y_{it-1}$, $\Delta X_{it} = X_{it} - X_{it-1}$ and $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$.

Note that in differencing the data, we lose the first time period for each individual: we now have $T-1$ observations for each i, rather than T. If we start with $T = 2$, then, after differencing, we arrive at one observation for each individual, and the transformed model (21) is just a cross-sectional model.

- The first-difference (FD) estimator of model (20) is defined as the pooled OLS estimator of the transformed model (21):

$$\begin{aligned} \hat{\beta}_{FD} &= \left(\sum_{i=1}^{N} \Delta X_i' \Delta X_i \right)^{-1} \sum_{i=1}^{N} \Delta X_i' \Delta Y_i \\ &= \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta X_{it}' \Delta X_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=2}^{T} \Delta X_{it}' \Delta y_{it} \end{aligned}$$

- For the FD estimator $\hat{\beta}_{FD}$ to be well defined, only time-varying (for at least some individuals) variables can actually be included in the explanatory variables X_{it}. Otherwise, ΔX_{it} would contain elements which are identically zero for all i and t. For the FD estimator $\hat{\beta}_{FD}$ to be well defined, the following assumption must thus holds:

  EC.3': In the sample (and thus in the population), each explanatory variable changes over time (for at least some i), and there is no exact linear relationship among them.

  In practice, this means that any time-constant variable originally included in the model – this includes the intercept – must be removed from X_i before differencing the data and computing the FD estimator[19]. As a result, only the parameter – and thus the partial effect – of the time-varying variables of the original model can actually be estimated. For more details, in particular regarding issues related to the presence of time dummies among the explanatory variables, see Wooldridge (2016), Section 13-5 (and also 14-1).

- Under assumptions EC.1, EC.2 and EC.3', the transformed model (21) satisfies the generic POLS assumptions POLS.1 – POLS.3 (linearity in parameters, random sampling, no perfect colinearity). Under assumption EC.4a, from (18), we

[19] All time-constant variables originally included in the model are 'differenced away' along with the individual effects when differencing the data. Formally, they are thus here simply treated as left in – be part of – the unobserved invididual effects.

have $E(\varepsilon_i|X_i) = 0$, so that :

$$E(\Delta\varepsilon_i|X_i) = E(D\varepsilon_i|X_i) = DE(\varepsilon_i|X_i) = D \cdot 0 = 0$$

and, as ΔX_i is a function of X_i, by the law of iterated expectations[20], we also have :

$$E(\Delta\varepsilon_i|\Delta X_i) = E\left[E(\Delta\varepsilon_i|X_i)|\Delta X_i\right] = E\left[0\,|\Delta X_i\right] = 0$$

In words, if in addition to assumptions EC.1, EC.2 and EC.3', assumption EC.4a also holds, then the transformed model (21) also satisfies the generic POLS assumption POLS.4 (zero conditional mean). Further, if in addition the conditional variance assumption EC.5a holds, from (19), we have $V(\varepsilon_i|X_i) = E(\varepsilon_i\varepsilon_i'|X_i) = \sigma_\varepsilon^2 I_T$, so that :

$$\begin{aligned} V(\Delta\varepsilon_i|X_i) &= E(\Delta\varepsilon_i\Delta\varepsilon_i'|X_i) = E(D\varepsilon_i\varepsilon_i'D'|X_i) \\ &= DE(\varepsilon_i\varepsilon_i'|X_i)D' = D(\sigma_\varepsilon^2 I_T)D' = \sigma_\varepsilon^2 DD' \end{aligned}$$

and, as ΔX_i is a function of X_i, by the law of iterated expectations, we also have :

$$\begin{aligned} V(\Delta\varepsilon_i|\Delta X_i) &= E(\Delta\varepsilon_i\Delta\varepsilon_i'|\Delta X_i) = E\left[E(\Delta\varepsilon_i\Delta\varepsilon_i'|X_i)|\Delta X_i\right] \\ &= E\left[V(\Delta\varepsilon_i|X_i)|\Delta X_i\right] = E\left[\sigma_\varepsilon^2 DD'|\Delta X_i\right] = \sigma_\varepsilon^2 DD' \qquad (22) \end{aligned}$$

In words, as $V(\Delta\varepsilon_i|\Delta X_i) = \sigma_\varepsilon^2 DD' \neq \sigma_\varepsilon^2 I_{T-1}$, under assumption EC.5a, the transformed model (21) does not satisfy the generic POLS assumption POLS.5 (homoskedasticity and no serial correlation). For assumption POLS.5 to hold, assumption EC.5a must be replaced by the assumption :

EC.5a' : $V(\Delta\varepsilon_i|X_i) = \sigma_\varepsilon^2 I_{T-1}$, $i = 1, ..., N$

As a matter of fact, under assumption EC.5a', from (22), we have :

$$V(\Delta\varepsilon_i|\Delta X_i) = E\left[V(\Delta\varepsilon_i|X_i)|\Delta X_i\right] = E\left[\sigma_\varepsilon^2 I_{T-1}|\Delta X_i\right] = \sigma_\varepsilon^2 I_{T-1}$$

In detailed form, for each i, assumption EC.5a' must be read : $Var(\Delta\varepsilon_{it}|X_i) = \sigma_\varepsilon^2$, for all $t = 2, ..., T$, and $Cov(\Delta\varepsilon_{it}, \Delta\varepsilon_{is}|X_i) = 0$, for all $t \neq s$. It thus assumes both constant variance and no correlation across time of the first-differenced errors $\Delta\varepsilon_{it}$, conditional on X_i. Assumption EC.5a' holds if, conditional on X_i, ε_{it} follows a random walk[21], which is a rather strong assumption, but sometimes reasonable. Note that when $T = 2$, as the transformed model (21) is just a cross-sectional model, assumption EC.5a' actually only requires constant variance of the scalar errors $\Delta\varepsilon_i$, conditional on X_i.

- According to the above discussion, under assumptions EC.1, EC.2, EC.3' and EC.4a, the transformed model (21) satisfies the generic POLS assumptions POLS.1 – POLS.4 (linearity in parameters, random sampling, no perfect colinearity and zero conditional mean), and if assumption EC.5a' also holds, the

[20] As a reminder, in its most general form, the law of iterated expectations states that, if x is a function of w, then $E(y|x) = E\left[E(y|w)|x\right]$. See Wooldridge (2010) p. 19 for details.

[21] See Wooldridge (2016), Section 11-3a.

POLS assumption POLS.5 (homoskedasticity and no serial correlation) likewise holds. As a consequence, from the generic properties of the POLS estimator :

- Under assumptions EC.1, EC.2, EC.3' and EC.4a, the FD estimator $\hat{\beta}_{FD}$ is an unbiased and consistent estimator of β, and asymptotically valid – i.e., approximately valid for N sufficiently large – inference procedures are provided by the heteroskedasticity and autocorrelation robust inference procedures outlined in Section 2.1.2 - C computed after POLS estimation of the transformed model (21).

- If, in addition to assumptions EC.1, EC.2, EC.3' and EC.4a, assumption EC.5a' also holds, the FD estimator $\hat{\beta}_{FD}$ is not only an unbiased and consistent estimator of β, but also the best linear unbiased estimator (BLUE) of β, and the usual inference procedures computed after POLS estimation of the transformed model (21) are asymptotically valid, i.e., approximately valid for N sufficiently large. If, in addition, an appropriately stated normality assumption[22] EC.6a' holds, the usual inference procedures computed after POLS estimation of the transformed model (21) are not only asymptotically valid, but also exact in finite sample.

• Remarks :

- Modern econometric software provide panel data management tools which make easy to compute first-differenced data. In applied works, heteroskedasticity and autocorrelation robust standard errors and tests should systematically be considered, at least for comparison.

- Assumption EC.4a requires that the explanatory variables are strictly exogenous conditional on the individual effects. This is necessary for both the finite sample and the asymptotic properties of the FD estimator to hold, and essentially can not be relaxed[23]. As a reminder, this implies that neither feedback from the current value of the dependent variable to the future values of the explanatory variables nor lagged dependent variables are allowed.

4. Fixed effects estimation

• First differencing estimation is not the only way to consistently estimate the so-called fixed effects model, i.e., the error component model described in assumption EC.1 :

$$y_{it} = X_{it}\beta + a_i + \varepsilon_{it} \ , \ \ i = 1, ..., N, \ t = 1, ..., T$$

[22] EC.6a' : $\Delta\varepsilon_i | X_i \sim N(0, \sigma_\varepsilon^2 I_{T-1})$, $i = 1, ..., N$.

[23] In a nutshell, from Wooldridge (2010), Section 7.8, the essentially weakest assumption under which the POLS estimator of the transformed model (21) is consistent for β is that $E(\Delta X_{it}'\Delta\varepsilon_{it}) = 0$, for all $i = 1, ..., N$ and $t = 2, ..., T$. But $E(\Delta X_{it}'\Delta\varepsilon_{it}) = E\left[(X_{it}' - X_{it-1}')(\varepsilon_{it} - \varepsilon_{it-1})\right] = E(X_{it}'\varepsilon_{it}) - E(X_{it}'\varepsilon_{it-1}) - E(X_{it-1}'\varepsilon_{it}) + E(X_{it-1}'\varepsilon_{it-1})$, so that for having $E(\Delta X_{it}'\Delta\varepsilon_{it}) = 0$, it is not only needed that ε_{it} is uncorretaled with the contemporaneous explanatory variables X_{it}, but also with the past and future explanatory variables X_{it-1} and X_{it+1}, i.e., the explanatory variables must be strictly exogeneous (conditional on the individual effects).

$$\Leftrightarrow \quad Y_i = X_i\beta + a_i i_T + \varepsilon_i \ , \quad i = 1, ..., N \tag{23}$$

where the random sampling assumption EC.2 and the zero conditional mean assumption EC.4a:

$$E(\varepsilon_i|X_i, a_i) = 0 \ \Leftrightarrow \ E(Y_i|X_i, a_i) = X_i\beta + a_i i_T \ , \ i = 1, ..., N$$

are maintained, but nothing – i.e., no restriction – is assumed regarding the relationship between the (unobserved) individual effects a_i and the observed explanatory variables X_{it} (i.e., assumption EC.4b $E(a_i|X_i) = 0$ is not maintained).

- Another method, which is more popular and works better under certain assumptions, is fixed effects estimation. Instead of getting rid of the problematic individual effects a_i by first differencing the data, fixed effects estimation eliminates the unobserved effects a_i by time-demeaning the data. Let M stands for the $T \times T$ time-demeaning – or within – transformation matrix:

$$\begin{aligned} M &= I_T - i_T(i_T' i_T)^{-1} i_T' = I_T - \tfrac{1}{T} J_T \\ &= \begin{bmatrix} 1-\frac{1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & \cdots & -\frac{1}{T} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{T} & -\frac{1}{T} & \cdots & 1-\frac{1}{T} \end{bmatrix} \end{aligned} \tag{24}$$

Note that M is symmetrical (i.e., $M = M'$) and idempotent (i.e., $MM = M$). Because $Mi_T = 0$, premultiplying both sides of (23) by M yields the transformed model:

$$\ddot{Y}_i = \ddot{X}_i\beta + \ddot{\varepsilon}_i \ , \quad i = 1, ..., N \tag{25}$$

where $\ddot{Y}_i = MY_i$ and $\ddot{\varepsilon}_i = M\varepsilon_i$ are $T \times 1$ vectors equal to:

$$\ddot{Y}_i = \begin{bmatrix} \ddot{y}_{i1} \\ \vdots \\ \ddot{y}_{iT} \end{bmatrix} = \begin{bmatrix} y_{i1} - \bar{y}_i \\ \vdots \\ y_{iT} - \bar{y}_i \end{bmatrix} \quad \text{and} \quad \ddot{\varepsilon}_i = \begin{bmatrix} \ddot{\varepsilon}_{i1} \\ \vdots \\ \ddot{\varepsilon}_{iT} \end{bmatrix} = \begin{bmatrix} \varepsilon_{i1} - \bar{\varepsilon}_i \\ \vdots \\ \varepsilon_{iT} - \bar{\varepsilon}_i \end{bmatrix}$$

with $\bar{y}_i = \frac{1}{T}\sum_{t=1}^T y_{it}$ and $\bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^T \varepsilon_{it}$, and $\ddot{X}_i = MX_i$ is a $T \times k$ matrix equal to:

$$\ddot{X}_i = \begin{bmatrix} \ddot{X}_{i1} \\ \vdots \\ \ddot{X}_{iT} \end{bmatrix} = \begin{bmatrix} X_{i1} - \bar{X}_i \\ \vdots \\ X_{iT} - \bar{X}_i \end{bmatrix}$$

with likewise $\bar{X}_i = \frac{1}{T}\sum_{t=1}^T X_{it}$, i.e., in detailed form:

$$\ddot{y}_{it} = \ddot{X}_{it}\beta + \ddot{\varepsilon}_{it} \ , \quad i = 1, ..., N, \ t = 1, ..., T$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$, $\ddot{X}_{it} = X_{it} - \bar{X}_i$ and $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$.

Note that $\ddot{Y}_i$ is nothing but the $T \times 1$ vector of residuals from the OLS regression of y_{it} on a constant using the T observations of individual i.[24] Likewise, $\ddot{X}_i$ is nothing but the matrix of residuals from the OLS regressions of each explanatory variables x_{itj} $(j = 1, ..., k)$ on a constant using the T observations of individual i. Note also that no observation is lost after time-demeaning the data : we still have T observations for each individual i. But because by construction $\sum_{t=1}^{T} \ddot{y}_{it} = 0$ and $\sum_{t=1}^{T} \ddot{X}_{it} = 0$, these T observations are linearly dependent : one of the observations is redundant (for each i and for any variable, any observation is equal to minus the sum of the $(T-1)$ other observations).

- The fixed effects (FE) estimator – or within estimator – of model (23) is defined as the pooled OLS estimator of the transformed model (25) :

$$
\begin{aligned}
\hat{\beta}_{FE} &= \left(\sum_{i=1}^{N} \ddot{X}_i' \ddot{X}_i\right)^{-1} \sum_{i=1}^{N} \ddot{X}_i' \ddot{Y}_i \\
&= \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{X}_{it}' \ddot{X}_{it}\right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{X}_{it}' \ddot{Y}_i
\end{aligned}
$$

- As in the case of first differencing estimation, for the FE estimator $\hat{\beta}_{FE}$ to be well defined, only time-varying (for at least some individuals) variables can actually be included in the explanatory variables X_{it}. Otherwise, $\ddot{X}_{it}$ would contain elements which are identically zero for all i and t. For the FE estimator $\hat{\beta}_{FE}$ to be well defined, assumption EC.3' must thus likewise hold. In practice, this likewise means that any time-constant variable originally included in the model – this includes the intercept – must be removed from X_i before time-demeaning the data and computing the FE estimator[25], so that, as a result, only the parameter – and thus the partial effect – of the time-varying variables of the original model can likewise actually be estimated.

- The FE estimator $\hat{\beta}_{FE}$ may alternatively be obtained from the following so-called dummy variable regression model, where the individual effects a_i $(i = 1, ..., N)$ are treated as parameters to estimate along with β :

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} i_T & 0 & \cdots & 0 \\ 0 & i_T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & i_T \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} + \begin{bmatrix} \cdots X_1 \cdots \\ \cdots X_2 \cdots \\ \vdots \\ \cdots X_N \cdots \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}
$$

$$
\Leftrightarrow \quad Y = Wa + X\beta + \varepsilon \qquad (26)
$$

where, as already defined, Y is a $NT \times 1$ vector, X is a $NT \times k$ matrix and β is $k \times 1$ vector, ε is likewise a $NT \times 1$ vector, and W is a $NT \times N$ matrix containing a set of N individual dummies and a is a $N \times 1$ vector containing the individual effects treated as parameters to estimate. In more detailed form,

[24] As a reminder, the OLS estimator of the regression $y_{it} = \alpha + \nu_{it}, t = 1, ..., T \Leftrightarrow Y_i = i_T\alpha + \nu_i$, is given by $\hat{\alpha} = (i_T' i_T)^{-1} i_T' Y_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} = \bar{y}_i$.

[25] As in the first differencing case, all time-constant variables originally included in the model are 'time-demeaned away' along with the individual effects when time-demeaning the data. Formally, they are thus again simply treated as left in – be part of – the unobserved individual effects.

at the level of the individuals, model (26) can be written:

$$Y_i = W_i a + X_i\beta + \varepsilon_i \ , \quad i = 1, ..., N \tag{27}$$

where all elements have already been defined, except W_i which quite obviously is a $T \times N$ matrix with a $T \times 1$ vector i_T of ones in the i-th column and zeros elsewhere. By definition, because for each i, $W_i a = a_i i_T$, model (27) is identical to the fixed effects model (23). Simply, because the a_i are now treated as parameters to estimate along with β, it accordingly contains $N + k$ independent variables (the N individual dummies and the k explanatory variables[26]).

Let $\hat{\gamma}_{DV} = (\hat{a}', \hat{\beta}'_{DV})'$ denote the pooled OLS estimator of dummy variable regression model (26):

$$\hat{\gamma}_{DV} = \begin{bmatrix} \hat{a} \\ \hat{\beta}_{DV} \end{bmatrix} = (Q'Q)^{-1} Q'Y = \left(\sum_{i=1}^{N} Q_i' Q_i \right)^{-1} \sum_{i=1}^{N} Q_i' Y_i \tag{28}$$

where $Q = [W \ X]$ is a $NT \times (N + k)$ matrix, and accordingly, $Q_i = [W_i \ X_i]$ is a $T \times (N + k)$ matrix. Using algebra on partitioned regression[27], it may be shown[28] that $\hat{\beta}_{DV}$ and the i-th element $\hat{a}_i$ of the $N \times 1$ fixed effects estimator $\hat{a}$, are given by:

$$\hat{\beta}_{DV} = \left(\sum_{i=1}^{N} \ddot{X}_i' \ddot{X}_i \right)^{-1} \sum_{i=1}^{N} \ddot{X}_i' \ddot{Y}_i = \hat{\beta}_{FE} \tag{29}$$

$$\hat{a}_i = \bar{y}_i - \bar{X}_i \hat{\beta}_{FE} \tag{30}$$

In words, the FE estimator may equivalently be obtained as the pooled OLS estimator of the transformed model (25) or the pooled OLS estimator of the dummy variable regression model (26).

- Under assumptions EC.1, EC.2 and EC.3', the transformed model (25) satisfies the generic POLS assumptions POLS.1 – POLS.3 (linearity in parameters, random sampling, no perfect colinearity). Under assumption EC.4a, from (18), we have $E(\varepsilon_i|X_i) = 0$, so that:

$$E(\ddot{\varepsilon}_i|X_i) = E(M\varepsilon_i|X_i) = ME(\varepsilon_i|X_i) = M \cdot 0 = 0$$

and, as $\ddot{X}_i$ is a function of X_i, by the law of iterated expectations, we also have:

$$E(\ddot{\varepsilon}_i|\ddot{X}_i) = E[E(\ddot{\varepsilon}_i|X_i)|\ddot{X}_i] = E[\,0\,|\ddot{X}_i] = 0$$

In words, if in addition to assumptions EC.1, EC.2 and EC.3', assumption EC.4a also holds, then the transformed model (25) also satisfies the generic POLS assumption POLS.4 (zero conditional mean). As a consequence, from the generic properties of the POLS estimator, the FE estimator $\hat{\beta}_{FE}$ is thus an unbiased and consistent estimator of β, and asymptotically valid inference procedures are provided by the heteroskedasticity and autocorrelation robust inference procedures

[26] Which are all supposed time-varying (for at least some individuals), as assumed by assumption EC.3'.
[27] See the so-called Frisch-Waugh theorem in Wooldridge (2016), Appendix E-1a.
[28] For more details, see Wooldridge (2010), Section 10.5, and Matyas and Sevestre (2008), Chapter 2.

outlined in Section 2.1.2 - C computed after POLS estimation of the transformed model (25).

Further, if in addition to assumptions EC.1, EC.2, EC.3' and EC.4a, the conditional variance assumption EC.5a also holds, from (19), we have $V(\varepsilon_i|X_i) = E(\varepsilon_i\varepsilon_i'|X_i) = \sigma_\varepsilon^2 I_T$, so that[29] :

$$\begin{aligned} V(\ddot{\varepsilon}_i|X_i) &= E(\ddot{\varepsilon}_i\ddot{\varepsilon}_i'|X_i) = E(M\varepsilon_i\varepsilon_i'M'|X_i) \\ &= ME(\varepsilon_i\varepsilon_i'|X_i)M' = M(\sigma_\varepsilon^2 I_T)M' = \sigma_\varepsilon^2 MM' = \sigma_\varepsilon^2 M \end{aligned} \tag{31}$$

and, as $\ddot{X}_i$ is a function of X_i, by the law of iterated expectations, we also have :

$$\begin{aligned} V(\ddot{\varepsilon}_i|\ddot{X}_i) &= E(\ddot{\varepsilon}_i\ddot{\varepsilon}_i'|\ddot{X}_i) = E[E(\ddot{\varepsilon}_i\ddot{\varepsilon}_i'|X_i)|\ddot{X}_i] \\ &= E[V(\ddot{\varepsilon}_i|X_i)|\ddot{X}_i] = E[\sigma_\varepsilon^2 M|\ddot{X}_i] = \sigma_\varepsilon^2 M \end{aligned}$$

In words, as $V(\ddot{\varepsilon}_i|\ddot{X}_i) = \sigma_\varepsilon^2 M \neq \sigma_\varepsilon^2 I_T$, under assumption EC.5a, the transformed model (25) does not satisfy the generic POLS assumption POLS.5 (homoskedasticity and no serial correlation). As a result, we can not use Property 17 for arguing that, under assumptions EC.1, EC.2, EC.3', EC.4a and EC.5a, asymptotically valid inference procedures are provided by the usual inference procedures computed after POLS estimation of the transformed model (25).

However, because of the very special form[30] of $V(\ddot{\varepsilon}_i|\ddot{X}_i) = \sigma_\varepsilon^2 M$, it turns out that only a slight modification of these usual inference procedures is actually needed for making them asymptotically valid under assumptions EC.1, EC.2, EC.3', EC.4a and EC.5a. As a matter of fact, under assumptions EC.1, EC.2, EC.3' and EC.4a, the transformed model (25) satisfies the generic POLS assumptions POLS.1 – POLS.4, so that from Property 18 we have :

$$\sqrt{N}(\hat{\beta}_{FE} - \beta) \overset{d}{\longrightarrow} N(0, A^{-1}BA^{-1}) \tag{32}$$

where :

$$A = E(\ddot{X}_i'\ddot{X}_i) \quad \text{and} \quad B = E(\ddot{X}_i'\ddot{\varepsilon}_i\ddot{\varepsilon}_i'\ddot{X}_i)$$

As M is symmetrical and idempotent, we have :

$$\ddot{X}_i'\ddot{\varepsilon}_i = (MX_i)'M\varepsilon_i = X_i'M'M\varepsilon_i = X_i'M'\varepsilon_i = \ddot{X}_i'\varepsilon_i \tag{33}$$

On the other hand, if assumption EC.5a also holds, from (19), we have $V(\varepsilon_i|X_i) = E(\varepsilon_i\varepsilon_i'|X_i) = \sigma_\varepsilon^2 I_T$, so that, as $\ddot{X}_i$ is a function of X_i, by the law of iterated expectations, we also have :

$$E(\varepsilon_i\varepsilon_i'|\ddot{X}_i) = E\left[E(\varepsilon_i\varepsilon_i'|X_i)|\ddot{X}_i\right] = E\left[\sigma_\varepsilon^2 I_T|\ddot{X}_i\right] = \sigma_\varepsilon^2 I_T \tag{34}$$

[29] As a reminder, M is symmetrical (i.e., $M = M'$) and idempotent (i.e., $MM = M$).

[30] Note that the variance-covariance matrix $V(\ddot{\varepsilon}_i|\ddot{X}_i) = \sigma_\varepsilon^2 M$ is actually singular (i.e., not full rank : its rank is equal to $T - 1$). This is basically a consequence of the already outlined fact that, after time-demeaning, one of the T observations of each individual i is redundant.

From (33) and (34), by the law of iterated expectations, we thus have:

$$\begin{aligned} B &= E(\ddot{X}_i'\ddot{\varepsilon}_i\ddot{\varepsilon}_i'\ddot{X}_i) = E(\ddot{X}_i'\varepsilon_i\varepsilon_i'\ddot{X}_i) \\ &= E\left[E(\ddot{X}_i'\varepsilon_i\varepsilon_i'\ddot{X}_i|\ddot{X}_i)\right] = E\left[\ddot{X}_i'E(\varepsilon_i\varepsilon_i'|\ddot{X}_i)\ddot{X}_i\right] \\ &= E\left[\ddot{X}_i'(\sigma_\varepsilon^2 I_T)\ddot{X}_i\right] = \sigma_\varepsilon^2 E(\ddot{X}_i'\ddot{X}_i) = \sigma_\varepsilon^2 A \end{aligned}$$

so that, from (32), we finally have:

$$\sqrt{N}(\hat{\beta}_{FE} - \beta) \xrightarrow{d} N(0, \sigma_\varepsilon^2 A^{-1}) \tag{35}$$

or in terms of approximate finite sample distribution which can be used – when N is sufficiently large – for performing inference:

$$\hat{\beta}_{FE} \approx N(\beta, \sigma_\varepsilon^2 A^{-1}/N) \tag{36}$$

Although the transformed model (25) does not satisfy assumption POLS.5, because of the very special form of $V(\ddot{\varepsilon}_i|\ddot{X}_i) = \sigma_\varepsilon^2 M$, the distributional result (35), or equivalently the approximate distributional result (36), is still the same as the distributional result (8), or equivalently the approximate distributional result (11), on which is based the usual POLS inference procedures. For these usual POLS inference procedures to be valid, it is however not enough to have the distributional result (35), or equivalently the approximate distributional result (36). It is also necessary that consistent estimators of σ_ε^2 and A are used to estimate the asymptotic variance $Avar(\hat{\beta}_{FE}) = \sigma_\varepsilon^2 A^{-1}/N$. This is where there is a catch: the usual POLS inference procedures computed after POLS estimation of the transformed model (25) use as an estimator of the asymptotic variance $Avar(\hat{\beta}_{FE}) = \sigma_\varepsilon^2 A^{-1}/N$:

$$\hat{V}(\hat{\beta}_{FE}) = \hat{s}^2 \left(\tfrac{1}{N}\sum_{i=1}^{N}\ddot{X}_i'\ddot{X}_i\right)^{-1} /N = \hat{s}^2(\ddot{X}'\ddot{X})^{-1} \tag{37}$$

where:

$$\hat{s}^2 = \frac{1}{NT-k}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\ddot{\varepsilon}}_{it}^2\,, \quad \widehat{\ddot{\varepsilon}}_{it} = \ddot{y}_{it} - \ddot{X}_{it}\hat{\beta}_{FE} \quad \text{and} \quad \ddot{X} = \begin{bmatrix} \ddot{X}_1 \\ \vdots \\ \ddot{X}_N \end{bmatrix}$$

The catch is that $\hat{s}^2$ is not a consistent estimator of σ_ε^2. To see that $\hat{s}^2$ is not a consistent estimator of σ_ε^2, suppose that $\ddot{\varepsilon}_{it}$ is actually observed and consider the estimator:

$$\hat{s}^{*2} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\varepsilon}_{it}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{\varepsilon}_{it}^2\right)$$

From (31), we have $E(\ddot{\varepsilon}_{it}^2) = (1 - \frac{1}{T})\sigma_\varepsilon^2 = \frac{T-1}{T}\sigma_\varepsilon^2$, so that by the law of large numbers (LLN), we have:

$$\hat{s}^{*2} \xrightarrow{p} E\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{\varepsilon}_{it}^2\right) = \frac{1}{T}\sum_{t=1}^{T}E(\ddot{\varepsilon}_{it}^2) = \frac{T-1}{T}\sigma_\varepsilon^2$$

As $\widehat{\ddot{\varepsilon}}_{it}$ converges to $\ddot{\varepsilon}_{it}$ (because $\hat{\beta}_{FE} \xrightarrow{p} \beta$), and $\frac{NT}{NT-k} \to 1$ as $N \to \infty$, we also have $\hat{s}^2 \xrightarrow{p} \frac{T-1}{T}\sigma_\varepsilon^2 \neq \sigma_\varepsilon^2$.

An obvious fix to the problem is to replace $\hat{s}^2$ in (37) by an estimator which is consistent for σ_ε^2. Following the same reasoning as above, it is easily checked that such a consistent estimator of σ_ε^2 is given by:

$$\hat{s}_{FE}^2 = \frac{1}{N(T-1)-k} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\ddot{\varepsilon}}_{it}^2 \,, \quad \text{where } \widehat{\ddot{\varepsilon}}_{it} = \ddot{y}_{it} - \ddot{X}_{it}\hat{\beta}_{FE} \tag{38}$$

With the estimator $\hat{s}_{FE}^2$ replacing $\hat{s}^2$ in (37), so that the asymptotic variance $Avar(\hat{\beta}_{FE}) = \sigma_\varepsilon^2 A^{-1}/N$ is instead estimated using the estimator:

$$\hat{V}_{FE}(\hat{\beta}_{FE}) = \hat{s}_{FE}^2(\ddot{X}'\ddot{X})^{-1} \tag{39}$$

the usual POLS inference procedures computed after POLS estimation of the transformed model (25) are now asymptotically valid. We will hereafter refer to these modified inference procedures as the 'FE modified usual inference procedures'.

- Further properties of the FE estimator may be outlined by looking at $\hat{\beta}_{FE}$ as the pooled OLS estimator of the dummy variable regression model (26). As a matter of fact, under assumptions EC.1, EC.2, EC.3', EC.4a and EC.5a, the dummy variable regression model (26) – where the vector of individual effects a is treated as a vector of parameters to estimate and W is nonstochastic[31] – actually satisfies the classical linear model assumptions E.1 – E.4 stated in SLN-I[32]. Accordingly, under assumptions EC.1, EC.2, EC.3', EC.4a and EC.5a, from Property 1 and 2 in SLN-I, the pooled OLS estimator $\hat{\gamma}_{DV} = (\hat{a}', \hat{\beta}_{DV}')'$ defined in (28) is an unbiased estimator of $\gamma = (a', \beta')'$:

$$E\left(\hat{\gamma}_{DV}|X\right) = \gamma = \begin{bmatrix} a \\ \beta \end{bmatrix}$$

and its variance-covariance matrix is given by:

$$V\left(\hat{\gamma}_{DV}|X\right) = \sigma_\varepsilon^2 \left(Q'Q\right)^{-1}, \quad \text{where } Q = [W \; X]$$

Also, from Property 3 in SLN-I, $\hat{\gamma}_{DV}$ is the best linear unbiased estimator (BLUE) of $\gamma = (a', \beta')'$, and from Property 5 in SLN-I, an unbiased estimator of σ_ε^2 is given by:

$$\hat{s}_{DV}^2 = \frac{1}{NT-(N+k)} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\varepsilon}_{it}^2 \,, \quad \text{where } \widehat{\varepsilon}_{it} = y_{it} - \hat{a}_i - X_{it}\hat{\beta}_{DV}$$

As from (29) $\hat{\beta}_{DV} = \hat{\beta}_{FE}$, this implies that, under assumptions EC.1, EC.2, EC.3', EC.4a and EC.5a, the FE estimator $\hat{\beta}_{FE}$ is not only unbiased, but it is also the best linear unbiased estimator (BLUE) of β. Note that $\hat{a}$, and thus

[31] i.e., fixed in repeated sample, and may thus be treated as a matrix of constant.

[32] i.e., E.1 : $Y = Wa + X\beta + \varepsilon$, E.2 : rank($[W \; X]$) $= N + k$, E.3 : $E(\varepsilon|X) = 0 \Leftrightarrow E(Y|X) = Wa + X\beta$, and E.4 : $V(\varepsilon|X) = \sigma_\varepsilon^2 I_{NT} \Leftrightarrow V(Y|X) = \sigma_\varepsilon^2 I_{NT}$, where I_{NT} is a $NT \times NT$ identity matrix.

each $\hat{a}_i$ as defined in (30), is also the best linear unbiased estimator (BLUE) of a, but contrary to $\hat{\beta}_{FE}$, it is not consistent[33] with T fixed, as $N \to \infty$. Further, as $\hat{\beta}_{DV} = \hat{\beta}_{FE}$ and, from (30), $\hat{a}_i = \bar{y}_i - \bar{X}_i\hat{\beta}_{FE}$, we have:

$$\begin{aligned}\widehat{\varepsilon}_{it} &= y_{it} - \hat{a}_i - X_{it}\hat{\beta}_{DV} = y_{it} - (\bar{y}_i - \bar{X}_i\hat{\beta}_{FE}) - X_{it}\hat{\beta}_{FE} \\ &= (y_{it} - \bar{y}_i) - (X_{it} - \bar{X}_i)\hat{\beta}_{FE} = \ddot{y}_{it} - \ddot{X}_{it}\hat{\beta}_{FE} = \widehat{\ddot{\varepsilon}}_{it}\end{aligned}$$

so that, noting that $NT - (N + k) = N(T - 1) - k$, we also have:

$$\hat{s}^2_{DV} = \frac{1}{NT - (N + k)} \sum_{i=1}^{N}\sum_{t=1}^{T} \widehat{\varepsilon}^2_{it} = \frac{1}{N(T - 1) - k} \sum_{i=1}^{N}\sum_{t=1}^{T} \widehat{\ddot{\varepsilon}}^2_{it} = \hat{s}^2_{FE} \tag{40}$$

In words, the estimator $\hat{s}^2_{FE}$ that we suggested above in (38) as a replacement of $\hat{s}^2$ in order to make the usual inference procedures computed after POLS estimation of the transformed model (25) asymptotically valid is actually equal to $\hat{s}^2_{DV}$. This implies that $\hat{s}^2_{FE}$ is not only a consistent estimator of σ^2_ε, but also an unbiased estimator of σ^2_ε.

If in addition to assumptions EC.1, EC.2, EC.3', EC.4a and EC.5a, the following normality assumption also holds:

EC.6a: $\varepsilon_i|X_i, a_i \sim N(0, \sigma^2_\varepsilon I_T) \Leftrightarrow Y_i|X_i, a_i \sim N(X_i\beta + a_i i_T, \sigma^2_\varepsilon I_T)$, $i = 1, ..., N$

then the dummy variable regression model (26) actually satisfies the full set[34] of classical linear model assumptions E.1 – E.5, so that, from Property 4 in SLN-I, the pooled OLS estimator $\hat{\gamma}_{DV} = (\hat{a}', \hat{\beta}_{DV})'$ has an exact in finite sample distribution given by:

$$\hat{\gamma}_{DV}|X \sim N(\gamma, \sigma^2_\varepsilon (Q'Q)^{-1}), \text{ where } \gamma = (a', \beta')' \text{ and } Q = [W\ X]$$

and an unbiased estimator of its variance-covariance matrix $V(\hat{\gamma}_{DV}|X) = \sigma^2_\varepsilon$ $(Q'Q)^{-1}$ is given by:

$$\hat{V}_{DV}(\hat{\gamma}_{DV}) = \hat{s}^2_{DV}(Q'Q)^{-1}$$

Accordingly, all usual inference procedures – regarding both a and β – computed after POLS estimation of the dummy variable regression model (26) are valid and exact in finite sample.

Unsurprisingly, using algebra on partitioned regression[35], as $\hat{\beta}_{FE} = \hat{\beta}_{DV}$, it may

---

[33] Each time a new individual i is added, another unknown parameter a_i is added, so that information does not accumulate on a as $N \to \infty$. This situation is usually refered to as an 'incidental parameters problem'. By the way, $\hat{a}$ provides a practical exemple of an estimator which is unbiased but not consistent.

[34] i.e., assumptions E.1 – E.4 as defined above, and E.5: $\varepsilon|X \sim N(0, \sigma^2_\varepsilon I_{NT}) \Leftrightarrow Y|X \sim N(Wa + X\beta, \sigma^2_\varepsilon I_{NT})$.

[35] Using algebra on partitioned regression, it may be shown that $\hat{\beta}_{FE} = \hat{\beta}_{DV}$ has an exact in finite sample distribution given by:

$$\hat{\beta}_{FE}|X \sim N(\beta, \sigma^2_\varepsilon(\ddot{X}'\ddot{X})^{-1})$$

and that an unbiased estimator of its variance-covariance matrix is given by:

$$\hat{V}_{DV}(\hat{\beta}_{FE}) = \hat{s}^2_{DV}(\ddot{X}'\ddot{X})^{-1} = \hat{s}^2_{FE}(\ddot{X}'\ddot{X})^{-1} = \hat{V}_{FE}(\hat{\beta}_{FE})$$

For more details, see again Wooldridge (2010), Section 10.5, and Matyas and Sevestre (2008), Chapter 2.

be shown that regarding β, all these usual inference procedures computed after POLS estimation of the dummy variable regression model (26) are – provided that the same appropriate number of degrees of freedom $NT - (N + k) = N(T-1) - k$ (and not $NT - k$) is used – exactly the same as the FE modified usual inference procedures computed after POLS estimation of the transformed model (25) outlined above. This in particular means that the FE modified usual inference procedures computed after POLS estimation of the transformed model (25) are not only asymptotically valid, but also exact in finite sample if normality actually holds.

- The whole above discussion about the properties of the FE estimator may be summarized as follows:

  - Under assumptions EC.1, EC.2, EC.3' and EC.4a, the FE estimator $\hat{\beta}_{FE}$ is an unbiased and consistent estimator of β, and asymptotically valid – i.e., approximately valid for N sufficiently large – inference procedures are provided by the heteroskedasticity and autocorrelation robust inference procedures outlined in Section 2.1.2 - C computed after POLS estimation of the transformed model (25).

  - If, in addition to EC.1, EC.2, EC.3' and EC.4a, assumption EC.5a also holds, the FE estimator $\hat{\beta}_{FE}$ is not only an unbiased and consistent estimator of β, but also the best linear unbiased estimator (BLUE) of β, and the FE modified usual inference procedures – i.e., the usual procedures where $\hat{V}(\hat{\beta}_{FE})$ is replaced by $\hat{V}_{FE}(\hat{\beta}_{FE})$ – computed after POLS estimation of the transformed model (25) are asymptotically valid, i.e., approximately valid for N sufficiently large. If, in addition, the normality assumption EC.6a holds, the FE modified usual inference procedures computed after POLS estimation of the transformed model (25) are not only asymptotically valid, but also exact in finite sample[36].

- Remarks:

  - Modern econometric software provide a built-in fixed effects estimation routine, usually including options to compute robust standard errors and perform robust tests. In applied works, it is not uncommon that only the heteroskedasticity and autocorrelation robust standard errors and tests are considered and reported.

  - Assumption EC.4a requires that the explanatory variables are strictly exogenous conditional on the individual effects. As in the case of first differencing estimation, this is necessary for both the finite sample and the asymptotic properties of the FE estimator to hold, and essentially can not be relaxed[37]. As a reminder, this implies that neither feedback from the current value of the dependent variable to the future values of the

[36] provided that the appropriate number of degrees of freedom $N(T-1) - k$ (and not $NT - k$) is used.

[37] In a nutshell, from Wooldridge (2010), Section 7.8, the essentially weakest assumption under which the POLS estimator of the transformed model (25) is consistent for β is that $E(\ddot{X}_{it}'\ddot{\varepsilon}_{it}) = 0$, for all $i = 1, ..., N$ and $t = 1, ..., T$. But $E(\ddot{X}_{it}'\ddot{\varepsilon}_{it}) = E\left[(X_{it}' - \bar{X}_i')(\varepsilon_{it} - \bar{\varepsilon}_i)\right] = E(X_{it}'\varepsilon_{it}) - E(X_{it}'\bar{\varepsilon}_i) - E(\bar{X}_i'\varepsilon_{it}) + E(\bar{X}_i'\bar{\varepsilon}_i)$, so that for having $E(\ddot{X}_{it}'\ddot{\varepsilon}_{it}) = 0$, it is not only needed that ε_{it} is uncorretaled with the contem-

explanatory variables nor lagged dependent variables are allowed.

– In applied works, fixed effects estimation is more popular than first differencing estimation, for some good reasons (at least when $T > 2$)[38] : FE estimation is efficient (BLUE) under more appealing conditions (assumption EC.5a versus assumption EC5.a' for FD estimation), may be shown to be less sensitive to violation of the assumption of strict exogeneity of the explanatory variables, and most modern software provide built-in routines which makes it effortless to compute. FD estimation may however be preferable in some situations, in particular when the idiosyncratic errors are expected to be strongly positively serially correlated and/or when N is small and T large[39].

5. Random effect estimation

- We finally return to the estimation of the so-called random effects model, i.e., the error component model described in assumption EC.1 :

$$\begin{aligned} y_{it} &= X_{it}\beta + a_i + \varepsilon_{it} \ , \ \ i = 1, ..., N, \ t = 1, ..., T \\ \Leftrightarrow \ \ Y_i &= X_i\beta + a_i i_T + \varepsilon_i \ , \ \ i = 1, ..., N \end{aligned} \tag{41}$$

where the random sampling and the no perfect collinearity assumptions EC.2 and EC.3, the (full) zero conditional mean assumption EC.4 :

(a) $E(\varepsilon_i|X_i, a_i) = 0 \ \Leftrightarrow \ E(Y_i|X_i, a_i) = X_i\beta + a_i i_T \ , \ \ i = 1, ..., N$

(b) $E(a_i|X_i) = E(a_i) = 0 \ , \ \ i = 1, ..., N$

as well as the (full) conditional variance assumption EC.5 :

(a) $V(\varepsilon_i|X_i, a_i) = \sigma_\varepsilon^2 I_T \ \Leftrightarrow \ V(Y_i|X_i, a_i) = \sigma_\varepsilon^2 I_T \ , \ \ i = 1, ..., N$

(b) $Var(a_i|X_i) = \sigma_a^2 \ , \ \ i = 1, ..., N$

are assumed to hold. As a reminder, the most distinctive assumption of this model is assumption EC.4b, which states that the individual effects a_i are unrelated with – statistically, mean independent of – the observed explanatory variables (in all time periods) X_i.

- We saw in Section 2.1.3 that, letting u_i stand for the composite $u_i = a_i i_T + \varepsilon_i$, under assumption assumptions EC.1 – EC.5, we have :

$$E(u_i|X_i) = 0 \ , \ \ i = 1, ..., N$$

poraneous explanatory variables X_{it}, but also with the explanatory variables in all other periods, i.e., the explanatory variables must be strictly exogeneous (conditional on the individual effects).

[38] When $T = 2$, the FE estimator and the FD estimator may be shown to be identical (numerically equal).

[39] For more details, see Wooldridge (2016), Section 14-1b, and further, Wooldridge (2010), Section 10-7.

and

$$\begin{aligned} V(u_i|X_i) &= \sigma_a^2 J_T + \sigma_\varepsilon^2 I_T \\ &= \begin{bmatrix} \sigma_a^2 + \sigma_\varepsilon^2 & \sigma_a^2 & \cdots & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_\varepsilon^2 & \cdots & \sigma_a^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_a^2 & \sigma_a^2 & \cdots & \sigma_a^2 + \sigma_\varepsilon^2 \end{bmatrix}, \ i = 1, ..., N \end{aligned}$$

so that assumptions POLS.1 − POLS.4 (linearity in parameters, random sampling, no perfect colinearity and zero conditional mean) hold, but, by definition, the POLS assumption POLS.5 (homoskedasticity and no serial correlation) does not hold. As a consequence, the POLS estimator $\hat{\beta}_{POLS}$ is an unbiased and consistent estimator of β, but it is not the best linear unbiased estimator (BLUE) of β, i.e., it is not efficient.

- As for heteroskedasticity in the cross-sectional case, using the generalized least squares theory developed in Section 3.1 of SLN-III, a more efficient estimator, which properly takes into account the form of the conditional variance $V(u_i|X_i) = \sigma_a^2 J_T + \sigma_\varepsilon^2 I_T$, i.e., which properly accounts for the dependence in time of the errors u_{it} induced by the presence of the individual effects, may readily be obtained. Before proceeding, note that $V(u_i|X_i)$ can be written as:

$$V(u_i|X_i) = \sigma_a^2 J_T + \sigma_\varepsilon^2 I_T = \sigma_\varepsilon^2 (I_T + \tfrac{\sigma_a^2}{\sigma_\varepsilon^2} J_T) = \sigma_\varepsilon^2 \Sigma$$

where $\Sigma = I_T + \frac{\sigma_a^2}{\sigma_\varepsilon^2} J_T$.

- Under assumptions EC.1 − EC.4, the error component model (41) satisfies the classical linear model assumptions E.1 − E.3 stated in SLN-I. If in addition assumption EC.5 also holds, then the following special case of assumption E.4bis stated in SLN-III also holds:

$$\text{E.4" Error component: } V(u|X) = \sigma^2\Omega \Leftrightarrow V(Y|X) = \sigma^2\Omega$$

where $\sigma^2 = \sigma_\varepsilon^2$ and Ω is a block-diagonal $NT \times NT$ matrix with $T \times T$ block-diagonal elements equal to $\Sigma = I_T + \frac{\sigma_a^2}{\sigma_\varepsilon^2} J_T$:

$$\Omega = \begin{bmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma \end{bmatrix}$$

so that:

$$\Omega^{-1} = \begin{bmatrix} \Sigma^{-1} & 0 & \cdots & 0 \\ 0 & \Sigma^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{-1} \end{bmatrix} \quad \text{and} \quad \Omega^{-\frac{1}{2}} = \begin{bmatrix} \Sigma^{-\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \Sigma^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{-\frac{1}{2}} \end{bmatrix}$$

Note that the nullity of all off-diagonal elements in Ω just follows from the independence of the observations across individuals implied by the random sampling assumption EC.2.

- From Property 11 in SLN-III, under assumptions E.1 – E.3 and assumption E.4", and thus likewise under assumptions EC.1 – EC.5, the best linear unbiased estimator (BLUE) of β is given by the special case of the generalized least squares (GLS) estimator:

$$\begin{aligned}\hat{\beta}_{GLS-EC} &= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y \\ &= \left(\sum_{i=1}^{n}X_i'\Sigma^{-1}X_i\right)^{-1}\sum_{i=1}^{n}X_i'\Sigma^{-1}Y_i \qquad (42)\end{aligned}$$

which by definition is the pooled OLS estimator of the transformed model:

$$Y^* = X^*\beta + u^* \ , \quad \text{where } Y^* = \Omega^{-\frac{1}{2}}Y \text{ and } X^* = \Omega^{-\frac{1}{2}}X$$

i.e., in more detailed form, at the level of the individuals, the pooled OLS estimator of the transformed model:

$$Y_i^* = X_i^*\beta + u_i^* \ , \quad i = 1, ..., N \qquad (43)$$

where the transformed variables are:

$$Y_i^* = \Sigma^{-\frac{1}{2}}Y_i \text{ and } X_i^* = \Sigma^{-\frac{1}{2}}X \ , \quad i = 1, ..., N$$

It may be shown[40] that the transformation matrix $\Sigma^{-\frac{1}{2}}$ may be written as:

$$\begin{aligned}\Sigma^{-\frac{1}{2}} &= I_T - \theta\tfrac{1}{T}J_T \\ &= \begin{bmatrix} 1-\theta\frac{1}{T} & -\theta\frac{1}{T} & \cdots & -\theta\frac{1}{T} \\ -\theta\frac{1}{T} & 1-\theta\frac{1}{T} & \cdots & -\theta\frac{1}{T} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta\frac{1}{T} & -\theta\frac{1}{T} & \cdots & 1-\theta\frac{1}{T} \end{bmatrix} \qquad (44)\end{aligned}$$

where $\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_a^2}}$, so that we have:

$$Y_i^* = \begin{bmatrix} y_{i1}^* \\ \vdots \\ y_{iT}^* \end{bmatrix} = \begin{bmatrix} y_{i1} - \theta\bar{y}_i \\ \vdots \\ y_{iT} - \theta\bar{y}_i \end{bmatrix} \text{ and } X_i^* = \begin{bmatrix} X_{i1}^* \\ \vdots \\ X_{iT}^* \end{bmatrix} = \begin{bmatrix} X_{i1} - \theta\bar{X}_i \\ \vdots \\ X_{iT} - \theta\bar{X}_i \end{bmatrix}$$

In words, the transformation matrix $\Sigma^{-\frac{1}{2}}$ simply means quasi-time-demeaning the data, i.e., expressing each variable in deviation from a fraction θ – necessarily between 0 and 1 – of its individual-specific mean. Note that the GLS estimator $\hat{\beta}_{GLS-EC}$ contains as limiting cases both the FE estimator (when $\theta \to 1$) and the POLS estimator (when $\theta \to 0$). Note also that, contrary to the time-demeaning – or within – transformation (24), the quasi-time-demeaning transformation (44) allows for time-constant variables.

[40] See Wooldridge (2010), Section 10.7.2.

- According to the generalized least squares theory developed in Section 3.1 of SLN-III, under assumptions EC.1 – EC.5 – which imply assumptions E.1 – E.3 and assumption E.4" –, the transformed model (43) satisfies the corresponding generic POLS assumptions POLS.1 – POLS.5 (linearity in parameters, random sampling, no perfect colinearity, zero conditional mean, homoskedasticity and no serial correlation). As a result:

  – Under assumptions EC.1 – EC.5, the GLS estimator $\hat{\beta}_{GLS-EC}$ is not only an unbiased and consistent estimator of β, but also the best linear unbiased estimator (BLUE) of β, and the usual inference procedures computed after POLS estimation of the transformed model (43) are asymptotically valid, i.e., approximately valid for N sufficiently large. If, in addition, an appropriately stated normality assumption[41] EC.6 holds, the usual inference procedures computed after POLS estimation of the transformed model (43) are not only asymptotically valid, but also exact in finite sample.

  – Under assumptions EC.1 – EC.4, i.e., without assuming that the conditional variance assumption EC.5 is correct, the GLS estimator $\hat{\beta}_{GLS-EC}$ is still an unbiased – but no longer BLUE – and consistent estimator of β, and asymptotically valid – i.e., approximately valid for N sufficiently large – inference procedures are provided by the heteroskedasticity and autocorrelation robust inference procedures outlined in Section 2.1.2 - C computed after POLS estimation of the transformed model (43).

- The GLS estimator $\hat{\beta}_{GLS-EC}$ depends on the unknown parameter θ, which itself depends on the unknown parameters σ_a^2 and σ_ε^2. But these parameters can be estimated from the data. There are different ways to this. The easiest way is to first estimate model (41) by pooled OLS, and retrieve the POLS residuals $\hat{u}_{it} = y_{it} - X_{it}\hat{\beta}_{POLS}$. It may be shown that[42], based on these POLS residuals and under assumptions EC.1 – EC.5, consistent estimators of σ_a^2 and σ_ε^2 are given by:

$$\hat{\sigma}_a^2 = \frac{1}{NT(T-1)/2 - k} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{u}_{it}\hat{u}_{is} \quad \text{and} \quad \hat{\sigma}_\varepsilon^2 = \hat{\sigma}_u^2 - \hat{\sigma}_a^2$$

  where:

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

  The consistency of these estimators basically relies on the law of large numbers (LLN) and on the facts that $\sigma_u^2 = \sigma_a^2 + \sigma_\varepsilon^2$, $E(u_{it}u_{is}) = \sigma_a^2$ for all i and $s \neq t$, $E(u_{it}^2) = \sigma_u^2$ for all i and t, and that $\hat{u}_{it}$ converges to u_{it} (because $\hat{\beta}_{POLS} \xrightarrow{p} \beta$).

- The random effects (RE) estimator of model (41) is defined as the GLS estimator $\hat{\beta}_{GLS-EC}$ where the unknown parameters σ_a^2 and σ_ε^2 are replaced by any

[41] EC.6: (a) $\varepsilon_i | X_i, a_i \sim N(0, \sigma_\varepsilon^2 I_T) \Leftrightarrow Y_i | X_i, a_i \sim N(X_i\beta + a_i i_T, \sigma_\varepsilon^2 I_T)$, $i = 1, ..., N$
(b) $a_i | X_i \sim N(0, \sigma_a^2)$, $i = 1, ..., N$

[42] For details, see Wooldridge (2010) Section 10.4.1.

consistent estimators such as $\hat{\sigma}_a^2$ and $\hat{\sigma}_\varepsilon^2$:

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{n} X_i'\hat{\Sigma}^{-1}X_i\right)^{-1} \sum_{i=1}^{n} X_i'\hat{\Sigma}^{-1}Y_i \ , \ \ \text{where } \hat{\Sigma} = I_T + \tfrac{\hat{\sigma}_a^2}{\hat{\sigma}_\varepsilon^2} J_T \qquad (45)$$

which by definition is the pooled OLS estimator of the transformed model:

$$\check{Y}_i^* = \check{X}_i^*\beta + \check{u}_i^* \ , \qquad i = 1, ..., N \qquad (46)$$

where the transformed variables are:

$$\check{Y}_i^* = \hat{\Sigma}^{-\frac{1}{2}}Y_i \ \text{ and } \ \check{X}_i^* = \hat{\Sigma}^{-\frac{1}{2}}X \ , \qquad i = 1, ..., N$$

with:

$$\hat{\Sigma}^{-\frac{1}{2}} = I_T - \hat{\theta}\tfrac{1}{T}J_T \ \text{ and } \ \hat{\theta} = 1 - \sqrt{\tfrac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2 + T\hat{\sigma}_a^2}}$$

i.e.:

$$\check{Y}_i^* = \begin{bmatrix} \check{y}_{i1}^* \\ \vdots \\ \check{y}_{iT}^* \end{bmatrix} = \begin{bmatrix} y_{i1} - \hat{\theta}\bar{y}_i \\ \vdots \\ y_{iT} - \hat{\theta}\bar{y}_i \end{bmatrix} \ \text{ and } \ \check{X}_i^* = \begin{bmatrix} \check{X}_{i1}^* \\ \vdots \\ \check{X}_{iT}^* \end{bmatrix} = \begin{bmatrix} X_{i1} - \hat{\theta}\bar{X}_i \\ \vdots \\ X_{iT} - \hat{\theta}\bar{X}_i \end{bmatrix}$$

- It may be shown that the replacement of σ_a^2 and σ_ε^2 by any consistent estimators such as $\hat{\sigma}_a^2$ and $\hat{\sigma}_\varepsilon^2$, which turns the GLS estimator $\hat{\beta}_{GLS-EC}$ into a feasible estimator, does not change its asymptotic properties[43]. More specifically:
  - Under assumptions EC.1 – EC.5, the RE estimator $\hat{\beta}_{RE}$ is consistent for β and asymptotically more efficient than the POLS estimator[44], and all usual inference procedures computed after POLS estimation of the transformed model (46) are still asymptotically valid, i.e., approximately valid for N sufficiently large.
  - Likewise, under only assumptions EC.1 – EC.4, i.e., without assuming that the conditional variance assumption EC.5 is correct, both $\hat{\beta}_{RE}$ is still consistent for β and the heteroskedasticity and autocorrelation robust inference procedures outlined in Section 2.1.2 - C computed after POLS estimation of the transformed model (46) are still asymptotically valid, i.e., approximately valid for N sufficiently large.

- Remarks:
  - Due to the estimation of the unknown parameters σ_a^2 and σ_ε^2, the RE estimator $\hat{\beta}_{RE}$ has no longer exact in finite sample properties: it is not unbiased, and the usual inference procedures computed after POLS estimation of the transformed model (46) are not exact even if the normality assumption EC.6 is assumed to hold.
  - Modern econometric software provide a built-in random effects estimation

[43] For a discussion, see Wooldridge (2010), Section 10.4.1 (and Section 7.5.1).
[44] This is just the asymptotic analog of the BLUE property of the GLS estimator.

routine, usually including options to compute robust standard errors and perform robust tests, so that there is no need to carry out multiple steps to perform RE estimation. In applied works, heteroskedasticity and autocorrelation robust standard errors and tests should systematically be considered, at least for comparison.

– If the conditional variance assumption EC.5 is not correct – for example due to heteroskedasticity in a_i and/or ε_{it}, or serial correlation in ε_{it} –, although still consistent[45], the RE estimator is no longer necessarily more efficient than the POLS estimator. In practice however, because the dependence in time – captured by σ_a^2 – of the composite errors u_{it} induced by the presence of the individual effects is usually found large, resorting to RE estimation along with robust inference procedures will often yield a more precise estimate of β than ignoring the individual effects altogether and using POLS likewise along with robust inference procedures.

– Assumption EC.4a requires that the explanatory variables are strictly exogenous conditional on the individual effects. As in the case of first differencing and fixed effects estimation, this is necessary for the properties of the RE estimator to hold, and essentially can not be relaxed[46]. As a reminder, this implies that neither feedback from the current value of the dependent variable to the future values of the explanatory variables nor lagged dependent variables are allowed. If one wants to allow for feedback from the current value of the dependent variable to the future values of the explanatory variables (allowing for dependent variables requires more sophisticated estimation techniques[47]), then the POLS estimator must be used (see the remark made at the end of Section 2.1.3).

– In empirical works, random effect estimation is notably useful when interest lies in the estimation of the partial effect of time-constant variables, something which can not be done with fixed effect or first differencing estimation. Just as in the cross-sectional case, if the estimated partial effects are to be confidently interpreted as genuine causal effects, then as many as possible relevant control variables, and in particular here as many as possible relevant time-constant control variables (with FE or FD estimation, such controls are not needed), should be included among the explanatory variables. This hopefully will minimize possible omitted variable bias and make more plausible the assumption that the (remaining) individual effects are unrelated with the observed explanatory variables.

– The assumption EC.4b that the individual effects are unrelated the observed explanatory variables may in practice be readily tested through a

[45] provided of course that assumptions EC.1 – EC.4 hold.

[46] In a nutshell, from Wooldridge (2010), Section 7.8, the essentially weakest assumption under which the POLS estimator of the transformed model (46) is consistent for β is that $E(X_{it}^{*\prime}u_{it}^{*}) = 0$, for all $i = 1, ..., N$ and $t = 1, ..., T$. But $E(X_{it}^{*\prime}u_{it}^{*}) = E\left[(X_{it}' - \theta\bar{X}_i')(u_{it} - \theta\bar{u}_i)\right] = E[(X_{it}' - \theta\bar{X}_i')(\varepsilon_{it} - \theta\bar{\varepsilon}_i + (1-\theta)a_i)] = E(X_{it}'\varepsilon_{it}) - \theta E(X_{it}'\bar{\varepsilon}_i) + (1-\theta)E(X_{it}'a_i) - \theta E(\bar{X}_i'\varepsilon_{it}) + \theta^2 E(\bar{X}_i'\bar{\varepsilon}_i) - \theta(1-\theta)E(\bar{X}_i'a_i)$, so that for having $E(X_{it}^{*\prime}u_{it}^{*}) = 0$, it is not only needed that ε_{it} is uncorretaled with the contemporaneous explanatory variables X_{it}, but also with the explanatory variables in all other periods (as well as that a_i is uncorretaled with the explanatory variables in all periods).

[47] See Wooldridge (2010), Section 11.6.2.

variable addition test. In a nutshell, this may simply be done by augmenting the considered model – which may contain both time-varying and time-constant variables – with the (individual) time-averages of its time-varying variables, and after RE estimation testing through a (preferably robust) F-test the nullity of the coefficients of the added time-average variables. See Wooldridge (2016), Section 14-3, for details. Note that this variable addition test is equivalent to the usual so-called Hausman test based on the direct comparison of RE and FE estimates (of the time-varying variables) of the considered model[48].

Reference

Matyas L. and P. Sevestre (2008), *The econometrics of Panel data*, Third Edition, Springer.

Wooldridge J.M. (2010), *Econometric Analysis of Cross-Section and Panel Data*, Second Edition, MIT Press.

Wooldridge J.M. (2016), *Introductory Econometrics : A Modern Approach*, 6th Edition, Cengage Learning.

[48] For more details, see Wooldridge (2010), Section 10.7.3.