The Econometrics of Multi-dimensional Panels

Chapter 1: Fixed Effects Models

version 1.7

Laszlo Balazsi, Laszlo Matyas, and Tom Wansbeek

Preliminary version, please do not quote

November 3, 2015

Introduction

Model formulations in which the individual and/or time heterogeneity factors are considered observable parameters, rather than random variables (see Chapter 2) are called fixed effects models. In the basic, most frequently used models, these heterogenous parameters are in fact splits of the regression constant. They can take different values in different sub-spaces of the original data space, while the slope parameters remain the same. This approach can then be extended to a varying coefficients framework, where heterogeneity is not picked up by the constant term, but rather by the slope coefficients.

In Section 1.1 we introduce the most frequently used models in a three-dimensional (3D) panel data setup. Section 1.2 deals with the Least Squares estimation of these models, while Section 1.3 analyses the behaviour of this estimator for incomplete/unbalanced data. Section 1.4 studies the properties of the so called Within estimator. Section 1.5 extends the original models to account for eventual heteroscedasticity and cross-correlation. Section 1.6 generalizes the presented models to four and higher dimensional data sets. Finally, Section 1.7 deals with the most relevant varying coefficients specifications.

1.1 Models with Different Types of Heterogeneity

In three-dimensional panel data, the dependent variable of a model is observed along three indices, such as y_{ijt}, $i = 1, \ldots, N_i$, $j = 1, \ldots, N_j$, and $t = 1, \ldots, T$, and the observations have the same ordering: index i goes the slowest, then j, and finally t the fastest,[1] such as

$$(y_{111} \cdots y_{11T} \cdots y_{1N_j1} \cdots y_{1N_jT} \cdots y_{N_i11} \cdots y_{N_i1T} \cdots y_{N_iN_j1} \cdots y_{N_iN_jT})' \,.$$

We assume in general that the index sets, $i \in \{1, \ldots, N_i\}$ and $j \in \{1, \ldots, N_j\}$ are (completely or partially) different. When dealing with economic flows, such as trade, capital, investment (FDI), etc., there is some kind of reciprocity, in such cases it is assumed, that $N_i = N_j = N$. The main question is how to formalize the individual and time heterogeneity — in our case, the fixed effects. In standard two–dimensional (2D) panels, there are only two effects, individual and time, so in principle 2^2 model specifications are possible (if we also count the model with no fixed effects). The situation is fundamentally different in three-dimensions. Strikingly, the 6 unique fixed effects formulations enable a great variety, precisely 2^6, of possible model specifications. Of course, only a subset of these are used, or make sense empirically, so in this chapter we are only considering the empirically most meaningful ones.

Throughout the chapter, we follow standard ANOVA notation, that is I and J denote the identity matrix, and the square matrix of ones respectively, with the size indicated in the index, $\bar{J}$ denotes the normalized J (each element is divided by the number in the index), and ι denotes the column vector of ones, with size in the index. Furthermore, an average over an index for a variable is indicated by a bar on the variable and a dot on the place of that index. When discussing unbalanced data, a plus sign at the place of an index indicates summation over that index. The matrix M with a subscript denotes projection orthogonal to the space spanned by the subscript.

[1] Please note, that the N_i, N_j notation does not mean, by itself, that the data is unbalanced.

The models can be casted in the general form

$$y = X\beta + D\pi + \varepsilon \tag{1.1}$$

with y and X being the vector and matrix of the dependent and explanatory variables (covariates) respectively of size $(N_iN_jT \times 1)$ and $(N_iN_jT \times K)$, β being the vector of the slope parameters of size $(K \times 1)$, π the composite fixed effects parameters, D the matrix of dummy variables, and finally, ε the vector of the disturbance terms.

The first attempt to properly extend the standard fixed effects panel data model (see, for example, Baltagi (2005) or Balestra and Krishnakumar (2008)) to a multidimensional setup was proposed by Matyas (1997). The specification of this model is

$$y_{ijt} = \beta' x_{ijt} + \alpha_i + \gamma_j + \lambda_t + \varepsilon_{ijt} \tag{1.2}$$

where the α_i, γ_j, and λ_t parameters are the individual and time-specific fixed effects (picking up the notation of (1.1), $\pi = (\alpha' \ \gamma' \ \lambda')'$), and ε_{ijt} are the $i.i.d.(0, \sigma_\varepsilon^2)$ idiosyncratic disturbance terms. We also assume that the x_{ijt} covariates and the disturbance terms are uncorrelated (this assumption is then relaxed in Chapter 3).

A model has been proposed by Egger and Pfaffermayr (2003), popular in the trade literature, which takes into account bilateral interaction effects. The model specification is

$$y_{ijt} = \beta' x_{ijt} + \gamma_{ij} + \varepsilon_{ijt}\,, \tag{1.3}$$

where the γ_{ij} are the bilateral specific fixed effect.

A variant of model (1.3), proposed by Cheng and Wall (2005), often used in empirical studies is

$$y_{ijt} = \beta' x_{ijt} + \gamma_{ij} + \lambda_t + \varepsilon_{ijt}\,. \tag{1.4}$$

It is worth noticing that models (1.3) and (1.4) are in fact straight 2D panel data models, where the individuals are now the (ij) pairs.

Baltagi et al. (2003), Baldwin and Taglioni (2006) and Baier and Bergstrand (2007) suggest several other forms of fixed effects. A simpler model is

$$y_{ijt} = \beta' x_{ijt} + \alpha_{jt} + \varepsilon_{ijt}\,, \tag{1.5}$$

where we allow the individual effect to vary over time. It is reasonable to present the symmetric version of this model (with α_{it} fixed effects); however, as it has the exact same properties, we consider the two models together.

A variation of this model is

$$y_{ijt} = \beta' x_{ijt} + \alpha_{it} + \alpha^*_{jt} + \varepsilon_{ijt}\,, \tag{1.6}$$

whereas the model that encompasses all the above effects is

$$y_{ijt} = \beta' x_{ijt} + \gamma_{ij} + \alpha_{it} + \alpha^*_{jt} + \varepsilon_{ijt}\,. \tag{1.7}$$

Table 1.1: Model specific D matrices

Model	D
(1.2)	$\big((I_{N_i} \otimes \iota_{N_jT}),\ (\iota_{N_i} \otimes I_{N_j} \otimes \iota_T),\ (\iota_{N_iN_j} \otimes I_T)\big)$
(1.3)	$\big(I_{N_iN_j} \otimes \iota_T\big)$
(1.4)	$\big((I_{N_iN_j} \otimes \iota_T),\ (\iota_{N_iN_j} \otimes I_T)\big)$
(1.5)	$\big(I_{N_i} \otimes \iota_{N_j} \otimes I_T\big)$
(1.6)	$\big((I_{N_i} \otimes \iota_{N_j} \otimes I_T),\ (\iota_{N_i} \otimes I_{N_jT})\big)$
(1.7)	$\big((I_{N_iN_j} \otimes \iota_T),\ (I_{N_i} \otimes \iota_{N_j} \otimes I_T),\ (\iota_{N_i} \otimes I_{N_jT})\big)$

Each model with its specific D matrix from formulation (1.1) is summarized in Table 1.1.

1.2 Least Squares Estimation of the Models

Let us assume, along with their independence from the disturbance terms, that the vector of regressors x_{ijt} is non-stochastic, and further, that none of the x_{ijt} variables is perfectly collinear with the fixed effects. In this case, if the matrix $(X,\ D)$ has full column rank, the Ordinary Least Squares (OLS) estimation of model (1.1), also called the Least Squares Dummy Variables (LSDV) estimator

$$\begin{pmatrix} \hat{\beta} \\ \hat{\pi} \end{pmatrix} = \begin{pmatrix} X'X & X'D \\ D'X & D'D \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ D'y \end{pmatrix},$$

is the Best Linear Unbiased Estimator (BLUE). This joint estimator, however, in some cases is cumbersome to implement, for example for model (1.3), as one has to invert a matrix of order $(K + N_i N_j)$, which can be quite difficult for large N_i and/or N_j. Nevertheless, following the Frisch-Waugh-Lovell theorem, or alternatively, applying partial inverse methods, the estimators can be expressed as

$$\begin{aligned} \hat{\beta} &= (X'M_D X)^{-1} X'M_D y \\ \hat{\pi} &= (D'D)^{-1} D'(y - X\hat{\beta}). \end{aligned} \tag{1.8}$$

where the idempotent and symmetric matrix $M_D = I - D(D'D)^{-1}D'$ is the so called *within projector*. In the usual panel data context, we call $\hat{\beta}$ in (1.8) the optimal Within estimator (due to its BLUE properties mentioned above). The LSDV estimator for β, $\hat{\beta}$, is different model-wise only to the extent of M_D. Table 1.2 captures these different projection matrices for all models discussed. Also, it is important to define the actual degrees of freedom to work with, so the third column of the table captures this. By using M_D, instead of possibly large matrices, we only have to invert a matrix of size $(K \times K)$ to get $\hat{\beta}$.

Table 1.2: Different forms of M_D after simplification

Model	M_D	Degrees of Freedom
(1.2)	$I-(I_{N_i}\otimes\bar{J}_{N_jT})-(\bar{J}_{N_i}\otimes I_{N_j}\otimes\bar{J}_T)-(\bar{J}_{N_iN_j}\otimes I_T)+2\bar{J}_{N_iN_jT}$	$N_iN_jT-N_i-N_j-T+1-K$
(1.3)	$I-(I_{N_iN_j}\otimes\bar{J}_T)$	$N_iN_j(T-1)-K$
(1.4)	$I-(I_{N_iN_j}\otimes\bar{J}_T)-(\bar{J}_{N_iN_j}\otimes I_T)+\bar{J}_{N_iN_jT}$	$(N_iN_j-1)(T-1)-K$
(1.5)	$I-(I_{N_i}\otimes\bar{J}_{N_j}\otimes I_T)$	$N_i(N_j-1)T-K$
(1.6)	$I-(I_{N_i}\otimes\bar{J}_{N_j}\otimes I_T)-(\bar{J}_{N_i}\otimes I_{N_jT})+(\bar{J}_{N_iN_j}\otimes I_T)$	$(N_i-1)(N_j-1)T-K$
(1.7)	$I-(I_{N_i}\otimes\bar{J}_{N_j}\otimes I_T)-(\bar{J}_{N_i}\otimes I_{N_jT})-(I_{N_iN_j}\otimes\bar{J}_T)+$ $(\bar{J}_{N_iN_j}\otimes I_T)+(\bar{J}_{N_i}\otimes I_{N_j}\otimes\bar{J}_T)+(I_{N_i}\otimes\bar{J}_{N_jT})-\bar{J}_{N_iN_jT}$	$(N_i-1)(N_j-1)(T-1)-K$

Estimation of the fixed effects parameters are captured by the second part of (1.8). Let us stop here to make an important remark. Notice, that so far we have assumed that D has full column rank. Unfortunately, this is only true for models of one fixed effect, that is, for (1.3) and (1.5). These model-specific estimators read as

$$\hat{\gamma}=\frac{1}{T}(I_{N_iN_j}\otimes\iota_T')(y-X\hat{\beta})$$

for model (1.3), and

$$\hat{\alpha}=\frac{1}{N_j}(I_{N_i}\otimes\iota_{N_j}'\otimes I_T)(y-X\hat{\beta})$$

for model (1.5). For the other models, the fixed effects are not identified. To make them identified, we have to use extra information in form of some restrictions over the fixed effects parameters. The two most widely used ones are either to normalize the fixed effects, or to leave the parameters belonging to the last (or first) individual or time period out. We will follow this latter approach. Let us illustrate the idea on model (1.2). For model (1.2), D has a rank deficiency of 2, but for the sake of symmetry, we leave out all three last fixed effects parameters, α_{N_i}, γ_{N_j}, and λ_T from the model, and add back a general constant term c. That is, for a given (ijt) observation $(i,j,t\neq N_i,N_j,T)$, the intercept is $c+\alpha_i+\gamma_j+\lambda_t$, but for example for $i=N_i$, it is only $c + \gamma_j + \lambda_t$. Let us denote this modified D dummy matrix by D^*, to stress that now it contains some extra information. As D^* has full column rank, estimator (1.8) works perfectly fine with D^*:

$$\hat{\pi}^* = (D^{*'} D^*)^{-1} D^{*'} (y - X\hat{\beta}) \,,$$

where now $\pi^* = (c', \ \alpha' \ \gamma' \ \lambda')'$. We may have a better understanding of these estimators, if we express them separately for each fixed effects parameters. This step, however, requires the introduction of complex matrix forms, and nontrivial manipulations, but as it turns out, using scalar notation, they can easily be represented. For model (1.2), this is

$$\begin{aligned}
\hat{c} \quad &= (\bar{y}_{N_i..} + \bar{y}_{.N_j.} + \bar{y}_{..T} - 2\bar{y}_{...}) - (\bar{x}'_{N_i..} + \bar{x}'_{.N_j.} + \bar{x}'_{..T} - 2\bar{x}'_{...})\hat{\beta} \\
\hat{\alpha}_i \quad &= (\bar{y}_{i..} - \bar{y}_{N_i..}) - (\bar{x}'_{i..} - \bar{x}'_{N_i..})\hat{\beta} \\
\hat{\gamma}_j \quad &= (\bar{y}_{.j.} - \bar{y}_{.N_j.}) - (\bar{x}'_{.j.} - \bar{x}'_{.N_j.})\hat{\beta} \\
\hat{\lambda}_t \quad &= (\bar{y}_{..t} - \bar{y}_{..T}) - (\bar{x}'_{..t} - \bar{x}'_{..T})\hat{\beta} \,.
\end{aligned}$$

Notice, that as we excluded α_{N_i} from the model, its estimator is indeed $\hat{\alpha}_{N_i} = (\bar{y}_{N_i..} - \bar{y}_{N_i..}) - (\bar{x}'_{N_i..} - \bar{x}'_{N_i..})\hat{\beta} = 0$, similarly for $\hat{\gamma}_{N_j}$, and $\hat{\lambda}_T$. For model (1.4),

$$\begin{aligned}
\hat{c} \quad &= (\bar{y}_{N_iN_j.} + \bar{y}_{..T} - \bar{y}_{...}) - (\bar{x}'_{N_iN_j.} + \bar{x}'_{..T} - \bar{x}'_{...})\hat{\beta} \\
\hat{\gamma}_{ij} \quad &= (\bar{y}_{ij.} - \bar{y}_{N_iN_j.}) - (\bar{x}'_{ij.} - \bar{x}'_{N_iN_j.})\hat{\beta} \\
\hat{\lambda}_t \quad &= (\bar{y}_{..t} - \bar{y}_{..T}) - (\bar{x}'_{..t} - \bar{x}'_{..T})\hat{\beta} \,.
\end{aligned}$$

For model (1.6), and (1.7), the rank deficiency, however, is not 2 but T, and $(N_i + N_j + T - 1)$, respectively. This means, that the restriction above can not be used. Instead, let us leave out the α_{it} parameters for $i = N_i$, that is, that last T from model (1.6). In this way, the estimators for the intercept parameters are

$$\begin{aligned}
\hat{\alpha}_{it} \quad &= (\bar{y}_{i.t} - \bar{y}_{N_i.t}) - (\bar{x}'_{i.t} - \bar{x}'_{N_i.t})\hat{\beta} \\
\hat{\alpha}^*_{jt} \quad &= (\bar{y}_{.jt} + \bar{y}_{N_i.T} - \bar{y}_{..t}) - (\bar{x}'_{.jt} + \bar{x}'_{N_i.T} - \bar{x}'_{..t})\hat{\beta} \,.
\end{aligned}$$

For model (1.7), we leave out γ_{ij} for $i = N_i$, α_{it} for $t = T$, and α^*_{jt} for $j = N_j$, and add back a general constant c. In this way, exactly $N_j + N_i + T - 1$ intercept parameters are eliminated, so the dummy matrix D^* has full rank. The estimators, with this D^* reads in scalar form

$$\begin{aligned}
\hat{c} &= (\bar{y}_{N_iN_j.} + \bar{y}_{N_i.T} + \bar{y}_{.N_jT} - \bar{y}_{N_i..} - \bar{y}_{.N_j.} - \bar{y}_{..T} + \bar{y}_{...}) \\
&\quad -(\bar{x}'_{N_iN_j.} + \bar{x}'_{N_i.T} + \bar{x}'_{.N_jT} - \bar{x}'_{N_i..} - \bar{x}'_{.N_j.} - \bar{x}'_{..T} + \bar{x}'_{...})\hat{\beta} \\
\bar{\gamma}_{ij} &= (\bar{y}_{ij.} - \bar{y}_{N_ij.} + \bar{y}_{i.T} - \bar{y}_{N_i.T} - \bar{y}_{i..} + \bar{y}_{N_i..}) \\
&\quad -(\bar{x}'_{ij.} - \bar{x}'_{N_ij.} + \bar{x}'_{i.T} - \bar{x}'_{N_i.T} - \bar{x}'_{i..} + \bar{x}'_{N_i..})\hat{\beta} \\
\bar{\alpha}_{it} &= (\bar{y}_{i.t} - \bar{y}_{i.T} + \bar{y}_{.N_jt} - \bar{y}_{.N_jT} - \bar{y}_{..t} + \bar{y}_{..T}) \\
&\quad -(\bar{x}'_{i.t} - \bar{x}'_{i.T} + \bar{x}'_{.N_jt} - \bar{x}'_{.N_jT} - \bar{x}'_{..t} + \bar{x}'_{..T})\hat{\beta} \\
\bar{\alpha}^*_{jt} &= (\bar{y}_{.jt} - \bar{y}_{.N_jt} + \bar{y}_{N_ij.} - \bar{y}_{N_iN_j.} - \bar{y}_{.j.} + \bar{y}_{.N_j.}) \\
&\quad -(\bar{x}'_{.jt} - \bar{x}'_{.N_jt} + \bar{x}'_{N_ij.} - \bar{x}'_{N_iN_j.} - \bar{x}'_{.j.} + \bar{x}'_{.N_j.})\hat{\beta}
\end{aligned}$$

Now, that we have derived appropriate estimators for all models, it is time to assess their properties. In finite samples, the OLS assumptions imposed guarantee that all estimators derived above are BLUE, with finite sample variances

$$V(\hat{\beta}) = \sigma^2_\varepsilon (X'M_D X)^{-1}$$

with the appropriate M_D, and

$$V(\hat{\pi}^*) = \sigma^2 (D^{*'}D^*)^{-1} + (D^{*'}D^*)^{-1} D^{*'} X V(\hat{\beta}) X' D^* (D^{*'}D^*)^{-1}$$

However, as multi-dimensional panel data are usually large in one or more directions, it is important to have a closer look at the asymptotic properties as well. Unlike cross-sectional, or time series data, panels can grow in multiple dimensions at the same time. As a matter of fact, three-way panel data can fall in one of the following seven asymptotic cases:

- $N_i \to \infty$, N_j, T fixed; $N_j \to \infty$, N_i, T fixed; $T \to \infty$, N_i, N_j fixed

- $N_i, N_j \to \infty$, T fixed; $N_i, T \to \infty$, N_j fixed; $N_j, T \to \infty$, N_i fixed
- $N_i, N_j, T \to \infty$.

It can be shown that $\hat{\beta}$ is consistent in all of the asymptotic cases for all models (if some weak properties hold). In order to make the models feasible for inference (i.e., for testing), we have to normalize the variances according to the asymptotics considered. When, for example, N_i goes to infinity, and N_j and T are fixed, $N_i V(\hat{\beta})$ is finite in the limit, as

$$\text{plim}_{N_i \to \infty} N_i V(\hat{\beta}) = \sigma_\varepsilon^2 \, \text{plim}_{N_i \to \infty} \left(\frac{X' M_D X}{N_i} \right)^{-1} = \sigma_\varepsilon^2 Q_{XMX}^{-1} \, ,$$

where Q_{XMX} is assumed to be a finite, positive semi-definite matrix. The estimators of fixed effects are consistent only if at least one of the indexes with which they are fixed with, is growing. For example, for model (1.2), $\hat{\alpha}_i$ is consistent only if N_j and/or T is going to infinity, and its variance is finite, and in addition, if it is pre-multiplied by N_j, in the case of $N_j \to \infty$, by T, in the case of $T \to \infty$, and by $N_j T$, when $N_j, T \to \infty$.

Testing for parameter values or restrictions is done in the usual way, using standard t-tests or F-tests. Typically, when one wants to test $\beta_k = 0$ or $\alpha_i = 0$, for example in model (1.2), the t-statistic is

$$\hat{\beta}_k / \sqrt{\hat{V}(\hat{\beta}_k)} \quad \text{and} \quad \hat{\alpha}_i / \sqrt{\hat{V}(\hat{\alpha}_i)} \, ,$$

where $V(\hat{\beta}_k)$ is the k-th diagonal element of $V(\hat{\beta})$, and $V(\alpha_i)$ is the diagonal element from $V(\hat{\pi}^*)$ corresponding to α_i. The degrees of freedom has to be adjusted accordingly, for each model, as Table 1.2 shows. The same degrees of freedom should be used when testing for the slope parameters and/or for the fixed effects of a given a model.

1.3 Some Data Problems

As in the case of the usual 2D panel data sets (see Wansbeek and Kapteyn (1989) or Baltagi (2005), for example), just more frequently, one may be faced with situations in which the data at hand is unbalanced. In our framework of analysis this means that $t \in T_{ij}$, for all (ij) pairs, where T_{ij} is a subset of the index set $t \in \{1, \ldots, T\}$, with T being chronologically the last time period in which we have any (i, j) observations. Note that two T_{ij} and $T_{i'j'}$ sets are usually different. A special case of incompleteness, which typically characterizes flow-type data, is the so-called no self-flow. In such data sets the individual index sets i and j are the same, so $N_i = N_j = N$ holds. Formally, this means, that, for all t, there are no observations when $i = j$, that is, we are missing a total NT of data points. In this section, however, we only consider general incompleteness, and take the no self-flow issue under lenses in Section 1.4.

In the case of incomplete data, the models can still be casted as in (1.1), but now D can not be represented nicely by kronecker products, as done in Table 1.1. However, with the incompleteness adjusted dummy matrices, $\tilde{D}$ (which we obtain from D by leaving out the rows corresponding to missing observations), the LSDV estimator of β and the fixed effects can still be worked out, maintaining its BLUE properties, following (1.8). There is, however, one practical obstacle in the way. Remember, that to reach $\hat{\beta}$ conveniently, we needed the exact form of M_D, which we collected for complete data in Table 1.2. As $\tilde{D}$ changes with different incompletenesses, $M_{\tilde{D}} = I - \tilde{D}(\tilde{D}'\tilde{D})^{-}\tilde{D}'$ can not be defined element-wise analytically, where "$-$" stands for any generalized inverse. Instead, we have to invert $(\tilde{D}'\tilde{D})$ directly, or use partitioned matrix inversion. Either way, we usually can not avoid large computational burdens when carrying out (1.8). Nevertheless, the estimators and the covariance matrices are obtained in the same way as for complete data (of course, after adjusting the matrices to incompleteness), and the properties of the estimators are the same as in the complete data case. Notice the crucial difference between $\tilde{D}$ and D^*: while $\tilde{D}$ has usually no full column rank, as we left out some rows from D (which also in general has no full column rank), D^* is simply designed to have full column rank (more precisely, to fix the rank deficiency in D). That is why we have to stick to generalized inverses for the former, but is enough to work with "simple" inverses for the latter dummy matrices.

Incompleteness is less of an issue in case of 2D models, where T is usually small, and N_i is large (that is when we have one high dimensional fixed effects), but is generally present in case of 3D data, where typically along with N_i, N_j is also large. In practice to alleviate this issue with the large dimensions, the best seems to be to turn to iterative solutions. One of the most widely used is based on the work of Carneiro et al. (2008), and later on Guimaraes and Portugal (2009). Let us show the procedure on model (1.2), the rest is a direct consequence. Model (1.2) in matrix form reads as

$$y = X\beta + \tilde{D}_1\alpha + \tilde{D}_2\gamma + \tilde{D}_3\lambda + \varepsilon\,, \tag{1.9}$$

where $\tilde{D}_k$ meant to stress, that the data is possibly incomplete: from the original $D_1 = (I_{N_i} \otimes \iota_{N_jT})$, $D_2 = (\iota_{N_i} \otimes I_{N_j} \otimes \iota_T)$, and $D_3 = (\iota_{N_iT} \otimes I_T)$, the rows matching with the missing observations are deleted. The normal equations from (1.9) are

$$\begin{aligned}
\beta &= (X'X)^{-1}X'(y - \tilde{D}_1\alpha - \tilde{D}_2\gamma - \tilde{D}_3\lambda) \\
\alpha &= (\tilde{D}_1'\tilde{D}_1)^{-}\tilde{D}_1'(y - X\beta - \tilde{D}_2\gamma - \tilde{D}_3\lambda) \\
\gamma &= (\tilde{D}_2'\tilde{D}_2)^{-}\tilde{D}_2'(y - X\beta - \tilde{D}_1\alpha - \tilde{D}_3\lambda) \\
\lambda &= (\tilde{D}_3'\tilde{D}_3)^{-}\tilde{D}_3'(y - X\beta - \tilde{D}_1\alpha - \tilde{D}_2\gamma)\,,
\end{aligned}$$

which suggests the so-called Gauss-Seidel, also called the "zigzag" algorithm, that is, we alternate between the estimation of β, and the fixed effects parameters, starting from some arbitrary initial values β^0, and $(\alpha^0, \gamma^0, \lambda^0)$. The computational improvement is clear: $(\tilde{D}_k'\tilde{D}_k)^{-}\tilde{D}_k$ defines a simple group average ($k = 1, 2, 3$) of the residuals, so the dimensionality issue is no longer a concern. Specifically, $(\tilde{D}_1'\tilde{D}_1)^{-}\tilde{D}_1'$ is translated into an average over (jt), $(\tilde{D}_2'\tilde{D}_2)^{-}\tilde{D}_2'$ an average over (it), and $(\tilde{D}_3'\tilde{D}_3)^{-}\tilde{D}_3'$ an average over (ij). Furthermore, $\tilde{D}_1\alpha$, etc. are just the columns of the current estimates of α, etc. After the sufficient number of steps, the iterative estimators all converge to the true LSDV.[2]

1.4 The Within Estimator

1.4.1 From the LSDV to the Within Estimator

As seen, the LSDV estimates all parameters of a model in one step. There is, however, an other appealing way to approach the estimation problem. The idea is that using orthogonal projections the slope parameters (and if needed the fixed effects) are estimated separately. First, with a projection orthogonal to D, we transform the model, in fact y and X, in such a way that clears the fixed effects. Then, we carry out an OLS estimation on the transformed variables $\tilde{y}$ and $\tilde{X}$. We have to point out, however, that unlike in the case of 2D models, there are usually multiple such Within transformations, which eliminate the fixed effects. Nevertheless, only the Within estimator

[2]The STATA program command *reg2hdfe* implements these results and can be found in the STATA Documentation. The code is designed to tackle two high dimensional fixed effects, however, it can be improved to treat three, or even more fixed effects at the same time.

based on the Within transformation originating from the LSDV conserves the BLUE properties and therefore is called the optimal one. To show this, notice, that as M_D is idempotent, the first part of formula (1.8) is equivalent to performing an OLS on

$$M_D y = M_D X\beta + \underbrace{M_D D}_{0}\pi + M_D\varepsilon\,,$$

where $M_D = I - D(D'D)^- D'$, as before. In the case of complete data, M_D can be translated into scalar notation, so we can fully avoid the dimensionality issue. Let us now go through all the models, and present the scalar form of the optimal Within transformation $M_D y$.

For model (1.2), the optimal transformation is

$$\tilde{y}_{ijt} = y_{ijt} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..t} + 2\bar{y}_{...} \tag{1.10}$$

As mentioned above, the uniqueness of the Within transformation is not guaranteed: for example transformation

$$\tilde{y}_{ijt} = y_{ijt} - \bar{y}_{ij.} - \bar{y}_{..t} + \bar{y}_{...}\,. \tag{1.11}$$

also eliminates the fixed effects from model (1.2). For model (1.3), the transformation is simply

$$\tilde{y}_{ijt} = y_{ijt} - \bar{y}_{ij.}\,. \tag{1.12}$$

For model (1.4), the optimal Within transformation is in fact (1.11). Notice, that model (1.2) is a special case of model (1.4) (with the restriction $\gamma_{ij} = \alpha_i + \gamma_j$), so while transformation (1.11) is optimal for (1.4), it is clear why it is not for the former: it "over-clears" the fixed effects, by not using the extra piece of information.

For model (1.5), the transformation is

$$\tilde{y}_{ijt} = y_{ijt} - \bar{y}_{.jt}\,, \tag{1.13}$$

while for models (1.6) and (1.7), they are

$$\tilde{y}_{ijt} = y_{ijt} - \bar{y}_{.jt} - \bar{y}_{i.t} + \bar{y}_{..t}\,, \tag{1.14}$$

and

$$\tilde{y}_{ijt} = y_{ijt} - \bar{y}_{ij.} - \bar{y}_{.jt} - \bar{y}_{i.t} + \bar{y}_{..t} + \bar{y}_{.j.} + \bar{y}_{i..} - \bar{y}_{...}\,, \tag{1.15}$$

respectively.

It can be seen, that the Within transformation work perfectly in wiping out the fixed effects. However, frequently in empirical applications, some explanatory variables, (i.e., some elements of the vector x'_{ijt}) do not span the whole (ijt) data space, that is, has some kind of "index deficiency". This means, that sometimes one (or more) of the regressors are perfectly collinear with one of the fixed effects. In such cases, we can consider that regressor as fixed, as it is wiped out along with the fixed effects. For example, for model (1.3), if we put an individual's gender among the regressors, $x_{ijt} \equiv x_i$ holds, and so is eliminated by the Within transformation (1.10). Clearly, parameters associated with such regressors then can not be estimated. This is most visible for model (1.8), as in that case all regressors fixed at least in one dimension are excluded from the model automatically after the Within transformation (1.15).

1.4.2 Incomplete Data with Within Estimators

We have covered briefly incompleteness in Section 1.3 already, but the Within estimators, and the underlying transformations, open a new way to deal with it. Let us start with the no self-flow data, and for a short time, assume, that the index sets i and j are the same, and so $N_i = N_j = N$.

In terms of the models from Section 1.1, the scalar transformations introduced there can no longer be applied. Fortunately, the pattern of the missing observations is highly structured, allowing for the derivation of optimal transformations that are still quite simple and maintain the BLUE properties of the Within estimators based on them. Following the derivations of Balazsi et al. (2015), the transformation for the models are the following:

$$\begin{aligned}\tilde{y}_{ijt} &= y_{ijt} - \tfrac{N-1}{N(N-2)T}(y_{i++} + y_{+j+}) - \tfrac{1}{N(N-2)T}(y_{j++} + y_{+i+}) \\ &\quad -\tfrac{1}{N(N-1)}y_{++t} + \tfrac{2}{N(N-2)T}y_{+++}\end{aligned} \tag{1.16}$$

for model (1.2), and

$$\tilde{y}_{ijt} = y_{ijt} - \frac{1}{T}y_{ij+} \tag{1.17}$$

for model (1.3). For models (1.4), and (1.5) the no self-flow transformations are

$$\tilde{y}_{ijt} = y_{ijt} - \frac{1}{T}y_{ij+} - \frac{1}{N(N-1)}y_{++t} + \frac{1}{TN(N-1)}y_{+++}\,, \tag{1.18}$$

and

$$\tilde{y}_{ijt} = y_{ijt} - \frac{1}{N-1}y_{+jt}\,, \tag{1.19}$$

while for models (1.6), and (1.7), they are

$$\begin{aligned}\tilde{y}_{ijt} &= y_{ijt} - \tfrac{N-1}{N(N-2)}\left(y_{i+t} + y_{+jt}\right) - \tfrac{1}{N(N-2)}\left(y_{+it} + y_{j+t}\right) \\ &\quad +\tfrac{1}{(N-1)(N-2)}y_{++t}\,,\end{aligned} \tag{1.20}$$

and

$$\begin{aligned}\tilde{y}_{ijt} = \;& y_{ijt} - \tfrac{N-3}{N(N-2)}(y_{i+t} + y_{+jt}) + \tfrac{N-3}{N(N-2)T}(y_{i++} + y_{+j+}) - \tfrac{1}{T}y_{ij+} \\ & +\tfrac{1}{N(N-2)}(y_{+it} + y_{j+t}) - \tfrac{1}{N(N-2)T}(y_{+i+} + y_{j++}) \\ & +\tfrac{N^2-6N+4}{N^2(N-1)(N-2)}(y_{++t} - y_{+++})\end{aligned} \tag{1.21}$$

respectively. So overall, the self-flow data problem can be overcome by using an appropriate Within transformation. Optimality of the estimators is preserved, as the transformations are derived from the Frisch-Waugh-Lovell theorem.

Next, let us go along these lines, and work out suitable Within transformations for any general form of incompleteness. Now, we are back in the case when i and j are different index sets. As the expressions below are all derived from the Frisch-Waugh-Lovell theorem, the transformations are optimal, and the estimators are BLUE. Remember, that now $t \in T_{ij}$, and let $R = \sum_{ij} |T_{ij}|$ denote the total number of observations, where $|T_{ij}|$ is the cardinality of the set T_{ij} (the number of observations in the given set).

For models (1.3) and (1.5), the unbalanced nature of the data does not cause any problem, the Within transformations can be used, and they have exactly the same properties, as in the balanced case. However, for models (1.2), (1.4), (1.6), and (1.7), we face some problems. As the Within transformations fail to fully eliminate the fixed effects for these models (somewhat similarly to the no self-flow case), the resulting Within estimators suffer from (potentially severe) biases. However, the Wansbeek and Kapteyn (1989) approach can be extended to these four cases.

Let us start with model (1.2). The dummy variable matrix D has to be modified to reflect the unbalanced nature of the data. Let the U_t and V_t ($t = 1 \dots T$) be the sequence of $(I_{N_i} \otimes \iota_{N_j})$ and $(\iota_{N_i} \otimes I_{N_j})$ matrices, respectively, in which the following adjustments are made: for each (ij) observation, we leave the row (representing (ij)) in U_t and V_t matrices untouched where $t \in T_{ij}$, but delete it from the remaining $T - |T_{ij}|$ matrices. In this way we end up with the following dummy variable setup

$$
\begin{aligned}
D_1^a &= (U_1', U_2', \dots, U_T')' \quad \text{of size} \quad (R \times N_i)\,, \\
D_2^a &= (V_1', V_2', \dots, V_T')' \quad \text{of size} \quad (R \times N_j)\,, \text{ and} \\
D_3^a &= \operatorname{diag}\{V_1 \cdot \iota_{N_i}, V_2 \cdot, \iota_{N_i} \dots, V_T \cdot \iota_{N_i}\} \quad \text{of size} \quad (R \times T)\,.
\end{aligned}
$$

So the complete dummy variable structure is now $D_a = (D_1^a, D_2^a, D_3^a)$. In this case, let us note here that, just as in Wansbeek and Kapteyn (1989), index t goes "slowly" and ij goes "fast". Using this modified dummy variable structure, the optimal projection removing the fixed effects can be obtained in three steps:

$$\begin{aligned} M_{D_a}^{(1)} &= I_R - D_1^a (D_1^{a\prime} D_1^a)^{-1} D_1^{a\prime}\,, \\ M_{D_a}^{(2)} &= M_{D_a}^{(1)} - M_{D_a}^{(1)} D_2^a (D_2^{a\prime} M_{D_a}^{(1)} D_2^a)^{-} D_2^{a\prime} M_{D_a}^{(1)}\,, \end{aligned}$$

and finally

$$M_{D_a} = M_{D_a}^{(3)} = M_{D_a}^{(2)} - M_{D_a}^{(2)} D_3^a (D_3^{a\prime} M_{D_a}^{(2)} D_3^a)^{-} D_3^{a\prime} M_{D_a}^{(2)}\,, \tag{1.22}$$

where "$-$" stands for any generalized inverse. It is easy to see that in fact $M_{D_a} D_a = 0$ projects out all three dummy matrices. Note that in the balanced case $(D_1^{a\prime} D_1^a)^{-1} = I_{N_i}/(N_j T)$, but now

$$(D_1^{a\prime} D_1^a)^{-1} = \text{diag}\left\{ \frac{1}{\sum_j |T_{1j}|}, \frac{1}{\sum_j |T_{2j}|}, \dots, \frac{1}{\sum_j |T_{N_i j}|} \right\} \quad \text{of size} \quad (N_i \times N_i)\,.$$

With this in hand, we only have to calculate two inverses instead of three, $(D_2^{a\prime} M_{D_a}^{(1)} D_2^a)^{-}$, and $(D_3^{a\prime} M_{D_a}^{(2)} D_3^a)^{-}$, with respective sizes $(N_j \times N_j)$ and $(T \times T)$. This is feasible for reasonable sample sizes.

For model (1.4), the job is essentially the same. Let the W_t $(t = 1 \dots T)$ be the sequence of $(I_{N_i N_j} \otimes I_{N_i N_j})$ matrices, where again for each (ij), we remove the rows corresponding to observation (ij) in those W_t, where $t \notin T_{ij}$. In this way,

$$\begin{aligned} D_1^b &= (W_1', W_2', \dots, W_T')' \quad \text{of size} \quad (R \times N_i N_j)\,, \\ D_2^b &= D_3^a \quad \text{of size} \quad (R \times T)\,. \end{aligned}$$

The first step in the projection is now

$$M_{D_b}^{(1)} = I_R - D_1^b (D_1^{b\prime} D_1^b)^{-1} D_1^{b\prime}\,,$$

so the optimal projection orthogonal to $D_b = (D_1^b, D_2^b)$ is simply

$$M_{D_b} = M_{D_b}^{(2)} = M_{D_b}^{(1)} - M_{D_b}^{(1)} D_2^b (D_2^{b\prime} M_{D_b}^{(1)} D_2^b)^{-} D_2^{b\prime} M_{D_b}^{(1)}\,. \tag{1.23}$$

Note that as

$$(D_1^{b\prime} D_1^b)^{-1} = \text{diag}\left\{\frac{1}{|T_{11}|}, \frac{1}{|T_{12}|}, \ldots, \frac{1}{|T_{N_iN_j}|}\right\} \quad \text{of size} \quad (N_iN_j \times N_iN_j)\,,$$

we only have to calculate the inverse of a $(T \times T)$ matrix – $D_2^{b\prime} M_{D_b}^{(1)} D_2^b$ – which is easily doable. Further, as discussed above, given that model (1.2) is nested in (1.4), transformation (1.23) is in fact also valid for model (1.2).

Let us move on to model (1.6). Now, after the same adjustments as before,

$$\begin{aligned} D_1^c &= \text{diag}\{U_1, U_2, \ldots, U_T\} \quad \text{of size} \quad (R \times N_iT) \quad \text{and} \\ D_2^c &= \text{diag}\{V_1, V_2, \ldots, V_T\} \quad \text{of size} \quad (R \times N_jT)\,, \end{aligned}$$

so the stepwise projection, removing $D_c = (D_1^c, D_2^c)$, is

$$M_{D_c}^{(1)} = I_R - D_1^c(D_1^{c\prime} D_1^c)^{-1} D_1^{c\prime}\,,$$

leading to

$$M_{D_c} = M_{D_c}^{(2)} = M_{D_c}^{(1)} - M_{D_c}^{(1)} D_2^c (D_2^{c\prime} M_{D_c}^{(1)} D_2^c)^{-} D_2^{c\prime} M_{D_c}^{(1)}\,. \tag{1.24}$$

Note that for M_{D_c}, we have to invert $(N_jT \times N_jT)$ matrices, which can be computationally difficult.

The last model to deal with is model (1.7). Let $D_d = (D_1^d, D_2^d, D_3^d)$, where the adjusted dummy matrices are all defined above:

$$\begin{aligned} D_1^d &= D_1^b \quad \text{of size} \quad (R \times N_iN_j)\,, \\ D_2^d &= D_1^c \quad \text{of size} \quad (R \times N_iT)\,, \\ D_3^d &= D_2^c \quad \text{of size} \quad (R \times N_jT)\,. \end{aligned}$$

Defining the partial projector matrices $M_{D_d}^{(1)}$ and $M_{D_d}^{(2)}$ as

$$\begin{aligned} M_{D_d}^{(1)} &= I_R - D_1^d(D_1^{d\prime} D_1^d)^{-1} D_1^{d\prime} \quad \text{and} \\ M_{D_d}^{(2)} &= M_{D_d}^{(1)} - M_{D_d}^{(1)} D_2^{d\prime} (D_2^{d\prime} M_{D_d}^{(1)} D_2^d)^{-} D_2^{d\prime} M_{D_d}^{(1)}\,, \end{aligned}$$

the appropriate transformation for model (1.7) is now

$$M_{D_d} = M_{D_d}^{(3)} = M_{D_d}^{(2)} - M_{D_d}^{(2)} D_3^{d'} (D_3^{d'} M_{D_d}^{(2)} D_3^d)^- D_3^{d'} M_{D_d}^{(2)} \,. \tag{1.25}$$

It can be easily verified that M_{D_d} is idempotent and $M_{D_d} D_d = 0$, so all the fixed effects are indeed eliminated.[3] As model (1.6) is covered by model (1.7), projection (1.25) eliminates the fixed effects from that model as well. Moreover, as suggested above, all three-way fixed effects models are in fact nested into model (1.7). It is therefore intuitive that transformation (1.25) clears the fixed effects in all model formulations. Using (1.7) is not always advantageous though, as the transformation involves the inversion of potentially large matrices (of order $N_i T$, and $N_j T$). In the case of most models studied, we can find suitable unbalanced transformations at the cost of only inverting $(T \times T)$ matrices; or in some cases, we can even derive scalar transformations. It is good to know, however, that there is a general projection that is universally applicable to all three-way models in the presence of all kinds of data issues.

It is worth noting that transformations (1.22), (1.23), (1.24), and (1.25) are all dealing in a natural way with the no self-flow problem, as only the rows corresponding to the $i = j$ observations need to be deleted from the corresponding dummy variable matrices.

All transformations detailed above can also be rewritten in a semi-scalar form. Let us show here how this idea works on transformation (1.25), as all subsequent transformations can be dealt with in the same way. Let

$$\phi = C^- \bar{D}' y \quad \text{and} \quad \omega = \tilde{C}^- (M_{D_d}^{(2)} D_3^d)' y \quad \xi = C^- \bar{D}' D_3^d \omega \,,$$

[3] A STATA program code for transformation (1.25) with a user-friendly detailed explanation is available at *www.personal.ceu.hu/staff/repec/pdf/stata-program_document-dofile.pdf.* Estimation of model (1.7) is then easily done for any kind of incompleteness.

where

$$C = \left(D_2^d\right)' \bar{D}\,, \quad \bar{D} = \left(I_R - D_1^d(D_1^{d'}D_1^d)^{-1}D_1^{d'}\right) D_2^d\,, \text{ and } \quad \tilde{C} = D_3^{d'} M_{D_d}^{(2)} D_3^d\,.$$

Now the scalar representation of transformation (1.25) is

$$\begin{aligned}[M_{D_d}y]_{ijt} = \;& y_{ijt} - \tfrac{1}{|T_{ij}|}\textstyle\sum_{t\in T_{ij}} y_{ijt} + \tfrac{1}{|T_{ij}|}a_{ij}'\phi - \phi_{it} \\ & -\omega_{jt} + \tfrac{1}{|T_{ij}|}\tilde{a}_{ij}'\omega + \xi_{it} - \tfrac{1}{|T_{ij}|}\left(a_{ij}^b\right)'\xi\,,\end{aligned}$$

where a_{ij} and $\tilde{a}_{ij}$ are the column vectors corresponding to observations (ij) from matrices $A = D_2^{d'}D_1^d$ and $\tilde{A} = D_3^{d'}D_1^d$, respectively; ϕ_{it} is the (it)-th element of the $(N_iT \times 1)$ column vector ϕ; ω_{jt} is the (jt)-th element of the $(N_jT \times 1)$ column vector ω; and finally, ξ_{it} is the element corresponding to the (it)-th observation from the $(N_iT \times 1)$ column vector, ξ.[4]

1.5 Heteroscedasticity and Cross-correlation

We have assumed so far throughout the chapter that the idiosyncratic disturbance terms in ε are in fact well behaved white noises, that is all heterogeneity is introduced into the model through the fixed effects. In some applications though this may be an unrealistic assumption, so next we relax it in two ways. We introduce heteroscedasticity and a simple form of cross correlation among the disturbance terms, and see how this influences the estimations methods introduced earlier. So far the approach has been to perform directly LSDV on the models, or alternatively, to transform the models in such a way that the fixed effects drop out, and then estimate the transformed models

[4]From a computational point of view, the calculation of matrix M_{D_d} is by far the most resource requiring as we have to invert $(N_iT \times N_iT)$, and $(N_jT \times N_jT)$ sized matrices. Simplifications related to this can dramatically reduce CPU and storage requirements. This topic, however, is well beyond the scope of this chapter.

with OLS. Now, however, in order to use all available information, the structure of the disturbances has to be taken into account for the estimation and use Feasible GLS (FGLS) instead of OLS on the fixed effects model. This can be done directly, expressing the partial estimators from the joint FGLS, or alternatively, after the appropriate transformation which removes the fixed effects, again, FGLS should be applied on the transformed model to take into account its covariance structure.

Formally, we project the data with some M_D orthogonal projection to get, with the familiar notation,

$$M_D y = M_D X\beta + \underbrace{M_D D}_{0}\pi + M_D\varepsilon \tag{1.26}$$

on which we perform (F)GLS to get the BLUE estimators:

$$\begin{aligned} \hat{\beta}_{(F)GLS} &= \left(X'M_D(M_D\Omega M_D)^{-1}M_DX\right)^{-1} X'M_D(M_D\Omega M_D)^{-1}M_Dy \\ \hat{\pi}_{(F)GLS} &= (D'D)^{-}D'(y - \hat{\beta}_{(F)GLS}X)\,, \end{aligned} \tag{1.27}$$

where $\Omega = E(\varepsilon\varepsilon')$. When ε follows a white noise process (what we have assumed so far), $\Omega = \sigma_\varepsilon^2 I$, and $\hat{\beta}_{(F)GLS} = (X'M_DX)^{-1} X'M_Dy$, is the OLS estimator on the transformed model. However, we can not make such simplification in general, when ε has some additional dependency embedded. To arrive to the (F)GLS estimators, first, we have to derive the covariance matrix of the model and analyze how the different transformations introduced earlier modify it. Then, we have to derive estimators for the variance components of the transformed model, in order to be able to use FGLS instead of OLS for the estimation.

1.5.1 The New Covariance Matrices

The initial assumptions about the disturbance terms are now replaced by

$$E\left(\varepsilon_{ijt}\varepsilon_{kls}\right)=\begin{cases}\sigma^2_{ij} & \text{if } i=k, j=l, t=s\\ \rho_1 & \text{if } i=k, j\neq l, \forall t,s\\ \rho_2 & \text{if } i\neq k, j=l, \forall t,s\\ 0 & \text{otherwise}\end{cases}$$

Then the variance-covariance matrix of all models introduced in Section 1.2 takes the form

$$E(\varepsilon\varepsilon')=\Omega=(\Upsilon\otimes I_T)+\rho_1(I_{N_i}\otimes J_{N_jT})+\rho_2(J_{N_i}\otimes I_{N_j}\otimes J_T), \qquad (1.28)$$

where

$$\Upsilon=\begin{pmatrix}\sigma^2_{11}-\rho_1-\rho_2 & 0 & \cdots & 0\\ 0 & \sigma^2_{12}-\rho_1-\rho_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2_{N_iN_j}-\rho_1-\rho_2\end{pmatrix}$$

is an $(N_iN_j\times N_iN_j)$ diagonal matrix. The transformed covariance matrix is obtained by pre- and post-multiplying Ω with the appropriate M_D,

$$E((M_D\varepsilon)(M_D\varepsilon)')=E(M_D\varepsilon\varepsilon' M_D)=M_D E(\varepsilon\varepsilon')M_D=M_D\Omega M_D\,.$$

Before presenting $M_D\Omega M_D$ for each model, let us make the following remark. From a practical point of view, ρ_1 and ρ_2 behaves as an i and j fixed effect, therefore if transformed properly, they are wiped out along with the other fixed effects. This is the case for models (1.2), (1.3), (1.4), (1.6), and (1.7), but not for model (1.5), as we will see later on. So for these five models,

$$M_D\Omega M_D=M_D(\Upsilon\otimes I_T)M_D\,.$$

One way to go forward is to elaborate on the above formula (by bringing in the specific M_D), however, we can rather use the idempotent property of M_D (that is, $M_D^- = M_D$, and $M_D^2 = M_D$) to first explore

$$(M_D(\Upsilon \otimes I_T)M_D)^{-1} = M_D(\Upsilon \otimes I_T)^{-1}M_D\,,$$

to get the (F)GLS of the transformed model (1.26) in turn as

$$\begin{aligned}\hat{\beta}_{(F)GLS} &= (X'M_D(M_D(\Upsilon \otimes I_T)M_D)^{-1}M_DX)^{-1} \times \\ &\quad \times X'M_D(M_D(\Upsilon \otimes I_T)M_D)^{-1}M_Dy \\ &= (X'M_D(\Upsilon \otimes I_T)^{-1}M_DX)^{-1}\,X'M_D(\Upsilon \otimes I_T)^{-1}M_Dy \\ &= (\tilde{X}'(\Upsilon \otimes I_T)^{-1}\tilde{X})^{-1}\tilde{X}'(\Upsilon \otimes I_T)^{-1}\tilde{y}\,.\end{aligned}$$

Notice, that now we only have to transform the data X and y (similarly as we did in Section 1.4), and use the same $(\Upsilon \otimes I_T)$ uniformly for all 5 models. As $(\Upsilon \otimes I_T)$ is a diagonal matrix, its inverse can be found at no cost (it is a diagonal matrix where the elements in the diagonal are the inverses of the elements from $(\Upsilon \otimes I_T)$).

Unfortunately, we are not that lucky with model (1.5), where the transformed covariance matrix, which we have to work with is, with $M_D = I - (I_{N_i} \otimes \bar{J}_{N_j} \otimes I_T)$, is

$$\begin{aligned}M_D\Omega M_D &= M_D(\Upsilon \otimes I_T)M_D + \rho_2 \cdot M_D(J_{N_i} \otimes I_{N_j} \otimes J_T)M_D \\ &= \left(I - (I_{N_i} \otimes \bar{J}_{N_j} \otimes I_T)\right)(\Upsilon \otimes I_T)\left(I - (I_{N_i} \otimes \bar{J}_{N_j} \otimes I_T)\right) \\ &\quad +\rho_2\left((J_{N_i} \otimes I_{N_j} \otimes J_T) - (J_{N_i} \otimes \bar{J}_{N_j} \otimes J_T)\right)\,.\end{aligned}$$

1.5.2 The Estimation of the Variance Components and the Cross Correlations

What now remains to be done is to estimate the variance components in order to make the GLS feasible. In principle, the job is to find a set of identifying equations from which the variance components can be expressed. Remember, that during the estimation we have transformed the models and performed an OLS on that. However this, in the case of some models, highly limits the number of identifying equations available for the variance components. For some models, this even means that the variance components are non-estimable without further restrictions on the structure of the disturbances (for example, $\rho_1 = \rho_2$, or an even stronger one, $\rho_1 = \rho_2 = 0$). This would certainly impede our cause, so let us take another track. Along with the OLS residuals from the transformed models, we can produce an other type of residual: the one from the LSDV estimation. As we will see, we can identify all the variance components from the LSDV residuals, and at the same time, we can obtain these residuals without the need to estimate the possibly numerous fixed effects.

As Section 1.2 suggests, whenever the D dummy coefficient matrix has no full column rank, the composite fixed effects parameters, π can not be identified (and of course, estimated). However, this is not the case for $D\pi$. The idea is, that for the LSDV residuals

$$\hat{\varepsilon} = y - X\hat{\beta} - D\hat{\pi} \tag{1.29}$$

we only need X, y, and $\hat{\beta}$, what we already have, and

$$D\hat{\pi} = D(D'D)^{-}D'(y - X\hat{\beta})\,.$$

Let us notice, that $D(D'D)^{-1}D' = I - M_D$, so in scalar terms, it defines a linear combination of sample means of $(y_{ijt} - x'_{ijt}\hat{\beta})$ for each model. If we denote $(y_{ijt} - x'_{ijt}\hat{\beta})$ by r_{ijt}, $D\hat{\pi}$ is obtained as in Table 1.3.

From this, having $\hat{\beta}$ and $D\hat{\pi}$ in hand, the LSDV residuals $\hat{\varepsilon}_{ijt}$ are obtained as in (1.29). With the residuals in hand, the variance components can

Table 1.3: Transformations on $r_{ijt} = (y_{ijt} - x'_{ijt}\hat{\beta})$ to get $D\hat{\pi}$

Model	$\tilde{r}_{ijt}$
(1.2)	$\tilde{r}_{ijt} = \bar{r}_{..t} + \bar{r}_{.j.} + \bar{r}_{i..} - 2\bar{r}_{...}$
(1.3)	$\tilde{r}_{ijt} = \bar{r}_{ij.}$
(1.4)	$\tilde{r}_{ijt} = \bar{r}_{..t} + \bar{r}_{ij.} - \bar{r}_{...}$
(1.5)	$\tilde{r}_{ijt} = \bar{r}_{.jt}$
(1.6)	$\tilde{r}_{ijt} = \bar{r}_{i.t} + \bar{r}_{.jt} - \bar{r}_{..t}$
(1.7)	$\tilde{r}_{ijt} = \bar{r}_{i.t} + \bar{r}_{.jt} + \bar{r}_{ij.} - \bar{r}_{i..} - \bar{r}_{.j.} - \bar{r}_{..t} + \bar{r}_{...}$

be expressed from the same identifying conditions regardless of the model specification. These are

$$\begin{aligned} E\left(\varepsilon_{ijt}^2\right) &= \sigma_{ij}^2 \\ E\left(\bar{\varepsilon}_{.jt}^2\right) &= \frac{1}{N_i^2}\left(\sum_i \sigma_{ij}^2 - N_i(N_i-1)\rho_2\right) \\ E\left(\bar{\varepsilon}_{i.t}^2\right) &= \frac{1}{N_j^2}\left(\sum_j \sigma_{ij}^2 - N_j(N_j-1)\rho_1\right) . \end{aligned}$$

The last step is to "estimate" the identifying conditions by replacing expectations with sample means, and the disturbances with the residuals. That is,

$$\begin{aligned} \hat{\sigma}_{ij}^2 &= \frac{1}{T}\sum_t \hat{\varepsilon}_{ijt}^2 \\ \hat{\rho}_2 &= \frac{1}{N_i(N_i-1)}\left[\sum_i \hat{\sigma}_{ij}^2 - \frac{1}{N_jT}\sum_{jt}\left(\sum_i \hat{\varepsilon}_{ijt}\right)^2\right] \\ \hat{\rho}_1 &= \frac{1}{N_j(N_j-1)}\left[\sum_j \hat{\sigma}_{ij}^2 - \frac{1}{N_iT}\sum_{it}\left(\sum_j \hat{\varepsilon}_{ijt}\right)^2\right] . \end{aligned} \tag{1.30}$$

Inserting these estimated variance components into (1.27) gives the FGLS estimator which handles the new and more flexible correlation structure.

When the data is incomplete, the above FGLS estimator is not appropriate as the Within transformation derived for complete data does not eliminate the fixed effects and, also, the covariance matrix Ω has to be adjusted to reflect the incomplete nature of the data, and so can not be represented with kronecker products. As the full analysis of such incomplete estimator would certainly be lengthy, we only provide some guidance on how to carry out the estimation. First, we have to transform the models in such a way, that the fixed effects drop out. This can be done according to the optimal incomplete-robust Within transformations derived in Section 1.4. Notice, that as the projection matrix M_D preserves the idempotent property in case of incomplete data as well (by construction), we only transform the y and x variables, and the transformed model still has the non-transformed, but now incompleteness adjusted covariance matrix. As the fixed effects are wiped out from the model, we can proceed by performing an FGLS estimation with the incomplete data adjusted covariance matrix $\tilde{\Omega}$, where $\tilde{\Omega}$ is obtained from the complete Ω by leaving out the rows and columns matching the missing observations. To get its inverse, $\tilde{\Omega}^{-}$ (where '$-$' is used to sign that Ω usually has no full rank when the data is incomplete), we have to use partial inverse methods, to avoid the dimensionality issue. The last step is to estimate the variance components, for which we only have to adjust (1.30) to the incomplete sample sizes. Note, that due the heteroscedasticity, the FGLS estimator is only consistent in T (T_{ij} in the incomplete case).

So how to proceed when the data is small in the time dimension? Let us consider the case of heteroscedasticity only, so set the cross correlations to zero ($\rho_1 = \rho_2 = 0$). This special case can be estimated in two ways. The optimal way is first to transform the model according to the optimal Within transformation as before, then carry out an FGLS with the heteroscedastic covariance matrix

$$\Omega_h = \text{diag}\left\{\sigma_{11}^2 I_{|T_{11}|},\ \sigma_{12}^2 I_{|T_{12}|},\ \ldots,\ \sigma_{N_i N_j}^2 I_{|T_{N_i N_j}|}\right\}\ ,$$

which is diagonal regardless of the potential data issues. The variance components are then estimated from

$$\hat{\sigma}^2_{ij} = \frac{1}{|T_{ij}|}\sum_t \hat{\varepsilon}^2_{ijt},$$

like before, with the $\hat{\varepsilon}_{ijt}$ being the LSDV residuals. However, this FGLS, as before, is still only T consistent. When the data is short in time, it is better to estimate the transformed model with OLS, which is still unbiased and consistent estimator of β in all the asymptotic cases studied before (though not optimal any more), and use heteroscedasticity robust White covariance matrix to estimate $V(\beta)$. Then we get

$$\begin{aligned} V(\hat{\beta}) &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\hat{\Omega}_h\tilde{X}(\tilde{X}'\tilde{X})^{-1} \\ &= \left(\sum_{ijt}\tilde{x}_{ijt}\tilde{x}'_{ijt}\right)^{-1}\left(\sum_{ijt}\tilde{x}_{ijt}\tilde{x}'_{ijt}\frac{1}{|T_{ij}|}\sum_t\hat{\varepsilon}^2_{ijt}\right)\left(\sum_{ijt}\tilde{x}_{ijt}\tilde{x}'_{ijt}\right)^{-1}, \end{aligned}$$

where '$\sim$' indicates that the variables are transformed. Notice again, that only the data X has to be transformed, but not Ω_h, due to the idempotent nature of the projections matrix.

1.6 Extensions to Higher Dimensions

In four and higher dimensions the number of specific effects, and therefore models, available is staggering. As a consequence, we have to restrict somehow the model formulations taken into account. The restriction used in this chapter is to allow for pairwise interaction effects only. Without attempting to be comprehensive, the most relevant four dimensional models are introduced in this section. Then, on a kind of benchmark model, we show intuitively how to estimate them for complete data, and also, in the case of the same data problems brought up in Sections 1.3 and 1.4. This is carried in a way that gives indications on how to proceed beyond four dimensions.

1.6.1 Different Forms of Heterogeneity

The dependent variable is now observed along four indexes, such as $ijst$. The generalization of model (1.4) (and also, that of the 2D fixed effects model with both individual and time effects) is

$$y_{ijst} = x'_{ijst}\beta + \gamma_{ijs} + \lambda_t + \varepsilon_{ijst} \, ,$$

or alternatively, a more restrictive formulation is

$$y_{ijst} = x'_{ijst}\beta + \alpha_i + \alpha^*_j + \gamma_s + \lambda_t + \varepsilon_{ijst} \, .$$

As in the case of 3D models, we can benefit from the multi-dimensional nature of the data, and let the fixed effects to be time dependent

$$y_{ijst} = x'_{ijst}\beta + \alpha_{it} + \gamma_{jt} + \delta_{st} + \varepsilon_{ijst}$$

that is we can allow all individual heterogeneity to vary over time as well. Finally, let us take the four-dimensional extension of the all-encompassing model (1.7), with pair-wise interaction effects:

$$y_{ijst} = x'_{ijst}\beta + \gamma^0_{ijs} + \gamma^1_{ijt} + \gamma^2_{jst} + \gamma^3_{ist} + \varepsilon_{ijst} \, , \tag{1.31}$$

with $i = 1 \dots N_i$, $j = 1 \dots N_j$, $s = 1 \dots N_s$, and $t = 1 \dots T$. This is what we consider from now on as a benchmark model, and show step-by-step how to estimate it.

1.6.2 Least Squares and the Within Estimators

If we keep maintaining the standard OLS assumptions lined up in Section 1.1, the LSDV estimator of model (1.31), following (1.8), is BLUE. Also, if we define the Within projector M_D, to get $\hat{\beta}$, the maximum matrix size to be worked with is still $(K \times K)$. For model (1.31), the composite dummy matrix D is

$$D = \big((I_{N_iN_jN_s} \otimes \iota_T),\ (I_{N_iN_j} \otimes \iota_{N_s} \otimes I_T),\ (\iota_{N_i} \otimes I_{N_jN_sT}),\ (I_{N_i} \otimes \iota_{N_j} \otimes I_{N_sT})\big)$$

with size $(N_iN_jN_sT \times (N_iN_jN_s + N_iN_jT + N_jN_sT + N_iN_sT))$ and column rank $(N_iN_jN_sT - (N_i - 1)(N_j - 1)(N_s - 1)(T - 1))$, leading to

$$\begin{aligned} M_D \;=\;& I_{N_iN_jN_sT} - \big(\bar{J}_{N_i} \otimes I_{N_jN_sT}\big) - \big(I_{N_i} \otimes \bar{J}_{N_j} \otimes I_{N_sT}\big) \\ & - \big(I_{N_iN_j} \otimes \bar{J}_{N_s} \otimes I_T\big) - \big(I_{N_iN_jN_s} \otimes \bar{J}_T\big) + \big(\bar{J}_{N_iN_j} \otimes I_{N_sT}\big) \\ & + \big(\bar{J}_{N_i} \otimes I_{N_j} \otimes \bar{J}_{N_s} \otimes I_T\big) + \big(\bar{J}_{N_i} \otimes I_{N_jN_s} \otimes \bar{J}_T\big) \\ & + \big(I_{N_i} \otimes \bar{J}_{N_jN_s} \otimes I_T\big) + \big(I_{N_i} \otimes \bar{J}_{N_j} \otimes I_{N_s} \otimes \bar{J}_T\big) + \big(I_{N_iN_j} \otimes \bar{J}_{N_sT}\big) \\ & - \big(\bar{J}_{N_iN_jN_s} \otimes I_T\big) - \big(\bar{J}_{N_iN_j} \otimes I_{N_s} \otimes \bar{J}_T\big) - \big(\bar{J}_{N_i} \otimes I_{N_j} \otimes \bar{J}_{N_sT}\big) \\ & - \big(I_{N_i} \otimes \bar{J}_{N_jN_sT}\big) + \bar{J}_{N_iN_jN_sT}\,. \end{aligned}$$

Just as before, M_D defines the optimal Within transformation to be performed on the data, so we can avoid matrix manipulations. That is, the LSDV estimator of β is analogous to the optimal Within estimator, which is obtained by first transforming the data according to

$$\begin{aligned} \tilde{y}_{ijst} \;=\;& y_{ijst} - \bar{y}_{.jst} - \bar{y}_{i.st} - \bar{y}_{ij.t} - \bar{y}_{ijs.} + \bar{y}_{..st} + \bar{y}_{.j.t} + \bar{y}_{.js.} \\ & + \bar{y}_{i..t} + \bar{y}_{i.s.} + \bar{y}_{ij..} - \bar{y}_{...t} - \bar{y}_{..s.} - \bar{y}_{.j..} - \bar{y}_{i...} + \bar{y}_{....} \end{aligned} \qquad (1.32)$$

(which eliminates $(\gamma^0_{ijs},\ \gamma^1_{ijt},\ \gamma^2_{jst},\ \gamma^3_{ist})$), then running an OLS on the transformed variables $\tilde{y}_{ijst}$, $\tilde{x}'_{ijst}$.

The properties of these estimators are identical to those of the three-way models, with the only modification, that now even more asymptotic cases could be considered. In general, the estimator of a fixed effects parameter is consistent, if an index with which the effect is fixed, goes to infinity. The resulting variances of any of the estimators should be normalized with the sample sizes which grow, and further, the degrees of freedom should be corrected to reflect the column rank deficiency in D.[5]

[5]For example for model (1.31), the correct degrees of freedom (coming from the rank

1.6.3 Some Data Problems

In theory, the missing data problem is corrected for by leaving out those rows from D which correspond to missing observations. LSDV estimation should then be done with the modified $\tilde{D}$, or alternatively, with $M_{\tilde{D}} = I - \tilde{D}(\tilde{D}'\tilde{D})^{-}\tilde{D}'$. Unfortunately, as now M_D has no clear structure, the resulting LSDV estimator can not be reached at an acceptable costs when the data is large. However, the optimal Within estimator offers a better way to tackle the problem at only moderate costs. Just like in Section 1.4, we have to come up with adjusted transformations, that clear out the fixed effects in the case of missing data. The no self-flow and unbalanced transformations in Section 1.4 can easily be generalized to any higher dimensions. For model (1.31), the no self-flow robust transformation reads as

$$\begin{aligned}\tilde{y}_{ijst} = \ & y_{ijst} - \tfrac{1}{N-1}y_{+jst} - \tfrac{1}{N-1}y_{i+st} - \tfrac{1}{N_s}y_{ij+t} - \tfrac{1}{T}y_{ijs+} + \tfrac{1}{(N-1)^2}y_{++st} \\ & + \tfrac{1}{(N-1)N_s}y_{+j+t} + \tfrac{1}{(N-1)T}y_{+js+} + \tfrac{1}{(N-1)N_s}y_{i++t} + \tfrac{1}{(N-1)T}y_{i+s+} \\ & + \tfrac{1}{N_sT}y_{ij++} - \tfrac{1}{(N-1)^2N_s}y_{+++t} - \tfrac{1}{(N-1)^2T}y_{++s+} - \tfrac{1}{(N-1)N_sT}y_{+j++} \\ & - \tfrac{1}{(N-1)N_sT}y_{i+++} + \tfrac{1}{(N-1)^2N_sT}y_{++++} - \tfrac{1}{(N-1)N_sT}y_{ji++} \\ & + \tfrac{1}{(N-1)T}y_{jis+} + \tfrac{1}{(N-1)N_s}y_{ji+t} - \tfrac{1}{N-1}y_{jist}\,. \end{aligned} \tag{1.33}$$

Note, that in the no self-flow case $N_i = N_j = N$ had to be assumed.

Incomplete data can also be handled quite flexibly in the case of four-dimensional models. Remember, that the key (iterative) unbalanced-robust transformation in Section 1.4 was (1.25), which can be generalized simply into a four dimensional setup. Let the dummy variables matrices for the four fixed effects in (1.31) be denoted by $D_e = (D_1^e, D_2^e, D_3^e, D_4^e)$ and let $M_{D_e}^{(k)}$ be the transformation that clears the first k fixed effects out; namely, $M_{D_e}^{(k)} \cdot (D_1^e, \dots, D_k^e) = (0, \dots, 0)$ for $k = 1 \dots 4$. The appropriate Within

of M_D) is $(N_i - 1)(N_j - 1)(N_s - 1)(T - 1) - K$.

transformation to clear out the first k fixed effects is then

$$M_{D_e}^{(k)} = M_{D_e}^{(k-1)} - \left(M_{D_e}^{(k-1)} D_k^e\right) \left[\left(M_{D_e}^{(k-1)} D_k^e\right)' \left(M_{D_e}^{(k-1)} D_k^e\right)\right]^{-} \left(M_{D_e}^{(k-1)} D_k^e\right)' , \tag{1.34}$$

where the first step in the iteration is

$$M_{D_e}^{(1)} = I - D_1^e \left((D_1^e)' D_1^e\right)^{-1} (D_1^e)' ,$$

and the iteration should be processed until $k = 4$. Note that none of this hinges on the model specification and can be done to any other multi-dimensional fixed effects model. The drawback, which can not really be addressed at this point, is again the increasing size of the matrices involved in the calculations. If this is the case, direct inverse calculations are feasible only up to some point, and further tricks (parallel computations, iterative inverting methods) should be used. However, this is beyond the scope of this chapter.

1.7 Varying Coefficients Models

So far we have assumed that the slope coefficients of the models considered are constant. This in fact meant that the heterogeneity was captured through the regression constant only, i.e., via the shifts of this term for different individuals and time points. One of the most important statistical feature of multidimensional data sets, however, is that heterogeneity is likely to take more complicated forms, which begs for more complex econometric models. One such approach with a more sophisticated form of heterogeneity is the varying coefficients model, where, along with the fixed effects, we allow the slope coefficients to vary as well.

The most general model we can imagine within this framework is

$$y_{ijt} = z'_{ijt}\gamma_{ijt} + \varepsilon_{ijt} \tag{1.35}$$

where we force some structure on γ_{ijt}. Note, that this is the general form of any standard multi-dimensional fixed effects model if we assume that $z'_{ijt} = (x'_{ijt},\ 1)$, and that $\gamma_{ijt} = (\beta',\ \pi'_{ijt})'$, with π_{ijt} being the composite fixed effect parameters.

The benchmark model we are focusing on, however, follows in spirit Balestra and Krishnakumar (2008) and Hsiao (1986), and takes the form

$$y_{ijt} = x'_{ijt}(\beta + \gamma_{ij} + \lambda_t) + \varepsilon_{ijt} \tag{1.36}$$

or similarly,

$$y = X_1\beta + X_2\gamma + X_3\lambda + \varepsilon$$

with

$$\begin{aligned} X_1 &\equiv \Delta(\iota_{N_iN_jT} \otimes I_K) && (N_iN_jT \times K) \\ X_2 &\equiv \Delta(I_{N_iN_j} \otimes \iota_T \otimes I_K) && (N_iN_jT \times N_iN_jK) \\ X_3 &\equiv \Delta(\iota_{N_iN_j} \otimes I_T \otimes I_K) && (N_iN_jT \times TK) \end{aligned}$$

where

$$\Delta = \begin{pmatrix} x'_{111} & & & \\ & x'_{112} & & \\ & & \ddots & \\ & & & x'_{N_iN_jT} \end{pmatrix} \quad (N_iN_jT \times N_iN_jTK)$$

is the diagonally arranged data matrix. Intuitively, this model suggests, that the explanatory variables have an effect on y through a common parameter β, but also through γ_{ij} and λ_t, which varies over individual pairs, and time periods. Notice, that $X = (X_1,\ X_2,\ X_3)$ has no full column rank, in fact has a rank deficiency of $2K$. Therefore, for identification $2K$ restrictions have to be imposed on the model. We can proceed by simply leaving out for example the last $2K$ of any parameters, however, a much less restrictive way, suggested by Hsiao (1986), is to normalize the average of the heterogeneous parameters:

$$\sum_{ij} \gamma_{ij} = 0\,; \quad \sum_{t} \lambda_t = 0\,. \tag{1.37}$$

Then $\tilde{X} = (X_1,\ \tilde{X}_2,\ \tilde{X}_3)$ has full column rank, where the $\tilde{X}_k$-s denote the matrices of observations X_k, after imposing the proper restrictions. To proceed, the adjusted model can be estimated with straight Least Squares optimally, to get

$$\left(\ \hat{\beta}' \quad \hat{\gamma}' \quad \hat{\lambda}'\ \right)' = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}y$$

or alternatively, partialling out γ and λ, and so expressing for β,

$$\hat{\beta} = \left(X_1' M_{\tilde{X}_2\tilde{X}_3} X_1\right)^{-1} X_1' M_{\tilde{X}_2\tilde{X}_3} y$$

with $M_{\tilde{X}_2\tilde{X}_3}$ being the projector matrix orthogonal to $(\tilde{X}_2,\ \tilde{X}_3)$. The problem is, that to get $M_{\tilde{X}_2\tilde{X}_3}$, we are faced of inverting $(KN_iN_j \times KN_iN_j)$ matrices, which becomes quickly computationally forbidding. One could try to figure out what this projection (with a set of non-trivial matrices) does to a typical x'_{ijt}, and be lost in the algebra quickly. Even if the above estimators can be computed for small samples, we still have the inconvenience of incorporating the restrictions first. Having said that, if we are uncertain about what the proper set of restriction would be, or simply there is scope for experimenting with different restrictions, we would have to redo the estimation each time.

There is, however, a more general, and useful approach to be used to derive estimators for β, and for the heterogeneous parameters as well. For this, we have to apply the theory of Least Squares of incomplete rank. Such estimators can be casted in the general form of

$$\hat{\delta} = \begin{pmatrix} \hat{\gamma} \\ \hat{\lambda} \\ \hat{\beta} \end{pmatrix} = \left(X'X\right)^{-} X'y + \left[\left(X'X\right)^{-} X'X - I\right] y_1 \tag{1.38}$$

with "–" standing for any generalized inverse, and as before, $X = (X_2,\ X_3,\ X_1)$ and y_1 being an arbitrary vector satisfying some regularity conditions.[6] To have an insight of the formula, notice that $\left(X'X\right)^{-} X'y$ is a general solution of the under-identified linear system of equations, while the second part incorporates the restriction needed for identification, through y_1. Detailed derivations are excluded here due to the lack of space, so we only present the final form of the estimators. With restrictions (1.37), the no full-rank estimator of model (1.36) is

$$\begin{aligned} \hat{\gamma}_{ij} &= \gamma^0_{ij} - \tfrac{1}{N_i N_j} \sum_{ij} \gamma^0_{ij} && (i, j = 1 \ldots N_i, N_j) \\ \hat{\lambda}_t &= \lambda^0_t - \tfrac{1}{T} \sum_t \lambda^0_t && (t = 1 \ldots T) \\ \hat{\beta} &= \beta^0 + \tfrac{1}{N_i N_j} \sum_{ij} \gamma^0_{ij} + \frac{1}{T} \sum_t \lambda^0_t \end{aligned} \tag{1.39}$$

where the "0" in the superscript stresses that those expressions are not yet estimators, but general solutions of the under-identified system of equation $(X'X)^{-}X'y$. Notice, that with the general solutions in hand (which we can get by repeatedly applying partial matrix operations), the estimators are obtained with simple "mean-corrections", following (1.39).

Fortunately, unbalanced data does not complicate our cause substantially, as the estimators are formulation wise equivalent to (1.39). Specifically, after we have found the general solutions β^0, γ^0 and λ^0 (in incomplete data), they can be used as in (1.39) to derive estimators.

[6] The reason for placing X_2 to the front of X is that $X_2'X_2$ is the largest matrix, yet block-diagonal. As its inverse is the inverses of its blocks, it is easily computed

As seen, this section only considered one specific model. Of course, there is substantial space for experimenting with other possible three-way specifications. For example, models

$$y_{ijt} = x'_{ijt}(\beta + \alpha_{it} + \alpha^*_{jt}) + \varepsilon$$

and

$$y_{ijt} = x'_{ijt}(\beta + \gamma_{ij} + \alpha_{it} + \alpha^*_{jt}) + \varepsilon$$

can also be considered, and can be estimated with the same steps and with slightly modified identifying restrictions as model (1.36). We must to keep track, however, of the total number of parameters to be estimated. For the last model considered, this number is $(1 + N_iN_j + N_iT + N_jT)K$ which can either be a classic case of over-specification, or in worse cases, can exceed the number of observations. This the main reason why this section focused on simpler models, like (1.36).

Naturally, nothing stops us from generalizing the above models to four, or even to higher dimensions, but computational requirements are frequently limiting the practical use of such formulations. The estimation of model

$$y_{ijst} = x'_{ijt}(\beta + \gamma_{ijs} + \lambda_t) + \varepsilon_{ijst}$$

has the same light computational requirement as model (1.36) (inverting a matrix of order T), but, for example, the estimation of

$$y_{ijst} = x'_{ijst}(\beta + \gamma^0_{ijs} + \gamma^1_{ijt} + \gamma^2_{jst} + \gamma^3_{ist}) + \varepsilon_{ijst}$$

involves matrices of order N_iN_jT, N_jN_sT, and N_iN_sT, which is forbidding even for moderate sample sizes.

References

Baier, Scott L. and Jeffrey H. Bergstrand, "Do Free Trade Agreements Actually Increase Members International Trade?," *Journal of International Economics*, 2007, *71*, 72–95.

Balazsi, Laszlo, Laszlo Matyas, and Tom Wansbeek, "The Estimation of Multi-dimensional Fixed Effects Panel Data Models," *Econometrics Review, forthcoming*, 2015.

Baldwin, Richard and Daria Taglioni, "Gravity for Dummies and Dummies for the Gravity Equations," *NBER Working Paper 12516*, 2006.

Balestra, Pietro and Jaya Krishnakumar, "Fixed Effects and Fixed Coefficients Models," in Matyas and Sevestre, eds., *The Econometrics of Panel Data, 3rd edition*, Springer Verlag, 2008, pp. 23–48.

Baltagi, Badi H., *Econometric Analysis of Panel Data, 3rd edition*, John Wiley & Sons, 2005.

_ , Peter Egger, and Michael Pfaffermayr, "A generalized Design for Bilateral Trade Flow Models.," *Economics Letters*, 2003, *80*, 391–397.

Carneiro, Anabela, Paulo Guimaraes, and Pedro Portugal, "Real Wages and the Business Cycle: Accounting for Worker and Firm Heterogeneity," *Unpublished manuscript*, 2008.

Cheng, I-Hui and Howard J. Wall, "Controlling for Heterogeneity in Gravity Models of Trade and Integration.," *Federal Reserve Bank of St. Louis Review*, 2005, *87*, 49–63.

Egger, Peter and Michael Pfaffermayr, "The Proper Econometric Specification of the Gravity Equation: 3-Way Model with Bilateral Interaction Effects.," *Empirical Economics*, 2003, *28*, 571–580.

Guimaraes, Paulo and Pedro Portugal, "A Simple Feasible Alternative Procedure to Estimate Models with High-Dimensional Fixed Effects," *IZA Discussion Paper No. 3935*, 2009.

Hsiao, Cheng, "Analysis of Panel Data," *Econometric Society Monographs, Cambridge University Press*, 1986.

Matyas, Laszlo, "Proper Econometric Specification of the Gravity Model.," *The World Economy*, 1997, *20*, 363–369.

Wansbeek, Tom and Arie Kapteyn, "Estimation of Error-Components Model With Incomplete Panels.," *Journal of Econometrics*, 1989, *41*, 341–361.