Panel Data Notes

Vasco Villas-Boas

April 20, 2025

CONTENTS

1 STATIC LINEAR PANEL DATA

There are two main motivations for the study of static panel data. First, we aim to correct for selection in cross-sectional datasets. That means that sampled individuals may have a component of the unobserved error that's *correlated* with one of the covariates. We try to estimate this individual specific component à la *fixed effects* to get consistent estimates for our parameters of interest.

Second, it can be that there's an individual specific unobserved component of the error that's *uncorrelated* with all of the covariates. In this case, we need to adjust the standard errors of our parameter estimates which is the focus of *random effects* models. In such settings we often aim to decompose the variability of an outcome. This model also allows us to estimate parameters on time-invariant regressors, which the fixed-effects model does not permit.

1.1 Fixed Effects Motivational Statistical Example

Consider the following population structural model with a scalar regressor where our goal is to identify β

$$y_{i1} = x_{i1}\beta + \eta_i + v_{i1},\ \mathbb{E}[v_{i1}|x_{i1}, \eta_i] = 0 \tag{1}$$

1.1.1 Observed Fixed Effect

If η_i is observed, then given a sample $\{(y_{i1}, x_{i1}, \eta_i)\}_{i=1}^N$, we can use OLS to estimate $(\hat{\beta}_N^O, \hat{\theta}_N^O)$ in the regression

$$y_{i1} = x_{i1}\beta + \theta\eta_i + v_{i1},\ \mathbb{E}[v_{i1}|x_{i1}, \eta_i] = 0$$

We will have that $(\beta, \theta) = \lim_{N\to\infty}(\hat{\beta}_N^O, \hat{\theta}_N^O)$ so that $\hat{\beta}_N^O$ is a consistent estimate of β.

1.1.2 Fixed Effect Uncorrelated to Covariate

Next suppose $\text{Cov}(x_{i1}, \eta_i) = 0$ and consider the following regression model where $u_{i1} = \eta_i + v_{i1}$.

$$y_{i1} = x_{i1}\beta + u_{i1},\ \mathbb{E}[u_{i1}|x_{i1}] = 0 \tag{2}$$

Note that the estimating assumption $\mathbb{E}[u_{i1}|x_{i1}] = 0$ holds by the assumption of this section (ie., $\text{Cov}(x_{i1}, \eta_i) = 0$) and the population assumption in equation (1). Given a sample $\{(y_{i1}, x_{i1})\}_{i=1}^N$, our estimate of β, $\hat{\beta}_N^{UC}$, from OLS in equation (2) has the following probability limit,

$$\begin{aligned}
\lim_{N\to\infty}\hat{\beta}_N^{UC} &= \lim_{N\to\infty}\frac{\hat{\text{Cov}}_N(y_{i1}, x_{i1})}{\hat{\text{Var}}_N(x_{i1})} \\
&= \lim_{N\to\infty}\frac{\hat{\text{Cov}}_N(x_{i1}\beta + u_{i1}, x_{i1})}{\hat{\text{Var}}_N(x_{i1})} \\
&= \lim_{N\to\infty}\frac{\hat{\text{Cov}}_N(x_{i1}\beta + \eta_i + v_{i1}, x_{i1})}{\hat{\text{Var}}_N(x_{i1})} \\
&= \lim_{N\to\infty}\frac{\beta\hat{\text{Cov}}_N(x_{i1}, x_{i1}) + \hat{\text{Cov}}_N(\eta_i, x_{i1}) + \hat{\text{Cov}}_N(v_{i1}, x_{i1})}{\hat{\text{Var}}_N(x_{i1})} \\
&= \lim_{N\to\infty}\beta + \frac{\hat{\text{Cov}}_N(\eta_i, x_{i1}) + \hat{\text{Cov}}_N(v_{i1}, x_{i1})}{\hat{\text{Var}}_N(x_{i1})} \\
&= \beta
\end{aligned}$$

The last step holds since $\lim_{N\to\infty}\hat{\text{Cov}}_N(\eta_i, x_{i1}) = \text{Cov}(\eta_i, x_{i1}) = 0$, $\lim_{N\to\infty}\hat{\text{Cov}}_N(v_{i1}, x_{i1}) = \text{Cov}(v_{i1}, x_{i1}) = 0$, $\lim_{N\to\infty}\hat{\text{Var}}_N(x_{i1}) = \text{Var}(x_{i1}) \in (0, \infty)$, by assumption, and then we apply the continuous mapping theorem.

1.1.3 Relevant Instrument Uncorrelated to Fixed Effect

There exists z_{i1} such that $\text{Cov}(z_{i1}, \eta_i) = 0$, $\text{Cov}(z_{i1}, v_{i1}) = 0$, and $\text{Cov}(x_{i1}, z_{i1}) \neq 0$. Then, we can do two stage least squares to get an estimate of β, $\hat{\beta}_N^{2SLS}$, in equation (1). We will have that

$$\begin{aligned}
\lim_{N\to\infty}\hat{\beta}_N^{2SLS} &= \lim_{N\to\infty}\frac{\hat{\text{Cov}}_N(y_{i1}, z_{i1})}{\hat{\text{Cov}}_N(x_{i1}, z_{i1})} \\
&= \lim_{N\to\infty}\frac{\hat{\text{Cov}}_N(x_{i1}\beta + \theta\eta_i + v_{i1}, z_{i1})}{\hat{\text{Cov}}_N(x_{i1}, z_{i1})} \\
&= \lim_{N\to\infty}\frac{\beta\hat{\text{Cov}}_N(x_{i1}, z_{i1}) + \hat{\text{Cov}}_N(\eta_i, z_{i1}) + \hat{\text{Cov}}_N(v_{i1}, z_{i1})}{\hat{\text{Cov}}_N(x_{i1}, z_{i1})} \\
&= \lim_{N\to\infty}\beta + \frac{\hat{\text{Cov}}_N(\eta_i, z_{i1}) + \hat{\text{Cov}}_N(v_{i1}, z_{i1})}{\hat{\text{Cov}}_N(x_{i1}, z_{i1})} \\
&= \beta
\end{aligned}$$

The last step holds since $\lim_{N\to\infty}\hat{\text{Cov}}_N(\eta_i, z_{i1}) = \text{Cov}(\eta_i, z_{i1}) = 0$, $\lim_{N\to\infty}\hat{\text{Cov}}_N(v_{i1}, x_{i1}) = \text{Cov}(v_{i1}, x_{i1}) = 0$, $\lim_{N\to\infty}\hat{\text{Cov}}_N(x_{i1}, z_{i1}) = \text{Cov}(x_{i1}, z_{i1}) \in \mathbb{R}$, by assumption, and then we apply the continuous mapping theorem.

1.1.4 First Differences

We have two samples from two time periods for each individual:

$$\begin{aligned}
y_{i1} &= x_{i1}\beta + \eta_i + v_{i1} \\
y_{i2} &= x_{i2}\beta + \eta_i + v_{i2} \\
\implies \Delta y_{i2} &= \Delta x_{i2}\beta + \Delta v_{i2}
\end{aligned}$$

where $\Delta r_{it} := r_{it} - r_{i(t-1)}$ for any random variable r and any $(i, t) \in \mathbb{N} \times \mathbb{N}$. Then, we can pose the following linear model:

$$\Delta y_{i2} = \Delta x_{i2}\beta + \Delta v_{i2},\ \mathbb{E}[\Delta v_{i2}|\Delta x_{i2}] = 0 \tag{3}$$

What's a sufficient condition for the estimating assumption of (3) to hold? We can assume that $\mathbb{E}[v_{it}|x_i, \eta_i] = 0\ \forall t \in \{1, 2\}$ where $x_i = [x_{i1}, x_{i2}]'$. If we also have that $\text{Var}(\Delta x_{i2}) \in (0, \infty)$, then, we can estimate β from this equation by $\hat{\beta}_N^{FD}$, which has probability limit

$$\begin{aligned}
\lim_{N\to\infty}\hat{\beta}_N^{FD} &= \lim_{N\to\infty}\frac{\hat{\text{Cov}}_N(\Delta y_{i2},\Delta x_{i2})}{\hat{\text{Var}}_N(\Delta x_{i2})} \\
&= \lim_{N\to\infty}\frac{\hat{\text{Cov}}_N(\Delta x_{i2}\beta+\Delta v_{i2},\Delta x_{i2})}{\hat{\text{Var}}_N(\Delta x_{i2})} \\
&= \lim_{N\to\infty}\frac{\beta\hat{\text{Cov}}_N(\Delta x_{i2},\Delta x_{i2})+\hat{\text{Cov}}_N(\Delta v_{i2},\Delta x_{i2})}{\hat{\text{Var}}_N(\Delta x_{i2})} \\
&= \lim_{N\to\infty}\beta+\frac{\hat{\text{Cov}}_N(\Delta v_{i2},\Delta x_{i2})}{\hat{\text{Var}}_N(\Delta x_{i2})} \\
&= \beta
\end{aligned}$$

The last step holds since $\lim_{N\to\infty}\hat{\text{Cov}}_N(\Delta v_{i2},\Delta x_{i2}) = \text{Cov}(\Delta v_{i2},\Delta x_{i2}) = 0$, $\lim_{N\to\infty}\hat{\text{Var}}_N(\Delta x_{i2}) = \text{Var}(\Delta x_{i2}) \in \mathbb{R}_{++}$, by assumption, and then we apply the continuous mapping theorem.

1.2 Classical Economic Example

Consider the following production function

$$y_{it} = x_{it}'\beta + \eta_i + v_{it},\ \mathbb{E}[v_{it}|x_{it},\eta_i] = 0$$

where y_{it} is log output of a single good by firm i at time t, x_{it} are log inputs of firm i at time t, η_i are inputs of firm i that remain constant over time such as soil quality, and v_{it} is a stochasitic input which is outside of the farmer's control that is unobserved by the econometrician. Note that we leave the joint distribution of (x_{it},η_i) completely unspecified. Let's additional posit that the farmer has profit and cost functions:

$$\Pi(x_{it},\eta_i,v_{it}),\ C(x_{it})$$

The farmer solves picks inputs to the point where expected marginal revenue equals marginal cost

$$\mathbb{E}[\Pi'(x_{it}^*,\eta_i,v_{it})|\eta_i,x_{it}^*] = C_l(x_{it}^*)$$

for each input good l. As econometricians, we observe $\{(\hat{y}_{it},\hat{x}_{it})\}_{(i,t)=(1,1)}^{(N,T)}$ and aim to estimate the parameter that linearly relates x_{it} to y_{it}.

1.3 Standard Static Fixed Effects Model

We assume that we have iid samples $\{(y_{i1},y_{i2},...,y_{iT},x_{i1},...,x_{iT})\}_{i=1}^N$ from the population equation

$$y_{it} = x_{it}'\beta + \eta_i + v_{it}$$

where we assume that N is large and T is relatively small compared to N and $x_{it}\in\mathbb{R}^K$. For notational convenience, I also define $x_i := [x_{i1},...,x_{iT}]$, $y_i := [y_{i1},...,y_{iT}]$, $v_i := [v_{i1},...,v_{iT}]$.

Assumption 1 (Strict Exogeneity)**.** $\mathbb{E}[v_i|x_i, \eta_i] = 0$

Note that Assumption (1) is equivalent to $\mathbb{E}[v_{it}|x_{i1}, ..., x_{iT}, \eta_i] = 0 \; \forall t \in \{1, ..., T\}$.

Assumption 2 (Homoskedasticity)**.** $\text{Var}(v_i|x_i, \eta_i) = \sigma^2 I_T$

Assumption 2 basically says that all $(v_{i1}, ..., v_{iT})$ are independent, have the same variance, and don't depend on time so that we have homoeskedasticity in the time series and in the cross-section.
These assumptions can be consolidated into the standard fixed effects model

Fixed Effects Regression Model

$$y_{it} = x_{it}'\beta + \eta_i + v_{it}$$
$$\mathbb{E}[v_i|x_i, \eta_i] = 0$$
$$\text{Var}(v_i|x_i, \eta_i) = \sigma^2 I_T$$
$$\text{for } v_i = [v_{i1}, ..., v_{iT}]', x_i = [x_{i1}, ..., x_{iT}]'$$

1.3.1 First Differences OLS

The first differences estimator, $\hat{\beta}_N^{FD}$, will be a consistent estimate of β as shown for the scalar regressor case in section (1.1.4).

For $T = 2$, we have that

$$\hat{\beta}_N^{FD} = \left(\sum_{i=1}^{N} \Delta x_{i2} \Delta x_{i2}' \right)^{-1} \left(\sum_{i=1}^{N} \Delta x_{i2} \Delta y_{i2} \right)$$

For $T \geq 3$, we have the $T - 1$ equations

$$\begin{aligned} \Delta y_{i2} &= \Delta x_{i2}'\beta + \Delta v_{i2} \\ \Delta y_{i3} &= \Delta x_{i3}'\beta + \Delta v_{i3} \\ &... \\ \Delta y_{iT} &= \Delta x_{iT}'\beta + \Delta v_{iT} \end{aligned}$$

It's useful to define the matrix $D := \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(T-1)\times T}$.

Then we can represent the $T - 1$ above equations[1] as $Dy_i = Dx_i + Dv_i$. By Assumption (1), $\mathbb{E}[Dv_i|x_i] = 0$ so that the following estimator $\hat{\beta}_N^{FD,OLS}$ is consistent for β.

[1]The $T = 2$ case also fits in this setting of $T - 1$ equations but I separate it above for didactic reasons. Coincidentally, we will also see that the $T = 2$ case is special in the sense that it's the only case when the first differences estimator is efficient.

$$\hat{\beta}_N^{FD,OLS} = \left(\sum_{i=1}^{N}(DX_i)'(DX_i)\right)^{-1}\left(\sum_{i=1}^{N}(DX_i)'(Dy_i)\right) \tag{4}$$

However, with $T \geq 3$, $\hat{\beta}_N^{FD,OLS}$ is not efficient. To see why, note that $\mathrm{Var}(Dv_i|X_i,\eta_i) = D\,\mathrm{Var}(v_i|X_i,\eta_i)D' = \sigma^2 DD'$. As we have seen in a previous statistics course, when the variance of the residuals is not a scaled identity matrix, then OLS is no longer efficient. It is GLS, which is efficient under Assumption (2).

1.3.2 Within Group GLS

The GLS estimator of β in this setting coincides with what we'll call the *within-group* estimator of β. We define

$$\hat{\beta}_N^{WG,GLS2} := \left(\sum_{i=1}^{N}(X_i'D')(DD')^{-1}(DX_i)\right)^{-1}\left(\sum_{i=1}^{N}(X_i'D')(DD')^{-1}Dy_i\right) \tag{5}$$

To get a better sense[2] of what transformation is happening in this GLS expression, let's compute $D'(DD')^{-1}D$.

Claim 1 (Within Group Transformation)**.** *D'(DD')$^{-1}$D* $= I_T - \frac{1}{T}1_T1_T'$

Proof.
Define $H := \begin{pmatrix} 1_T'T^{-1/2} \\ (DD')^{-1/2}D \end{pmatrix} \in \mathbb{R}^{T\times T}$. Note that $HH' = I_T$. Then, since H is square, that means that H' is the inverse of H so that $H'H = I_T$. Since $H'H = \frac{1}{T}1_T1_T' - D'(DD')^{-1}D$, then $D'(DD')^{-1}D = I_T - 1_T1_T'$. □

Now, let's get a proper sense of what $D'(DD')^{-1}D$ does to any vector r_i.

$$\begin{aligned} D'(DD')^{-1}Dr_i &= (I_T - \frac{1}{T}1_T1_T')r_i \\ &= r_i - \bar{r}_i 1_T \end{aligned}$$

where $\bar{r}_i = \frac{1}{T}\sum_{t=1}^{T} r_{it}$. Thus, the matrix $D'(DD')^{-1}D$ demeans all entries of the vector it multiplies.

Next, for brevity, define $Q := D'(DD')^{-1}D$. Note that Q is idempotent and symmetric.

$$\begin{aligned} QQ &= D'(DD')^{-1}DD'(DD')^{-1}D = D'(DD')^{-1}D \\ Q' &= (D'(DD')^{-1}D)' = D'(DD')^{-1}D = Q \end{aligned}$$

Thus, we can rewrite $\hat{\beta}_N^{WG,GLS2}$ as

$$\begin{aligned} \hat{\beta}_N^{WG,GLS2} &= \left(\sum_{i=1}^{N}X_i'Q'QX_i\right)^{-1}\left(\sum_{i=1}^{N}X_i'Q'Qy_i\right) \\ &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)\right) \end{aligned}$$

[2]Note that the '2' in the variable name of $\hat{\beta}_N^{WG,GLS2}$ refers to this estimator being the GLS estimator under Assumption (2).

Define $\tilde{r}_{it} = r_{it} - \bar{r}_i$ for any variable r, we can $\hat{\beta}_N^{WG,GLS2}$ as the OLS estimate of β for the linear regression model below:

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{v}_{it},\ \mathbb{E}[\tilde{v}_{it}|\tilde{x}_{it}] = 0 \tag{6}$$

Note that Assumption (1) is indeed sufficient for the estimating assumption of equation (6) to hold so that $\hat{\beta}_N^{WG,GLS2}$ is indeed consistent for β[3].

We can arrive at our estimate of $\hat{\beta}_N^{WG,GLS2}$ in a different manner as well. Consider stacking the cross-sectional entries as follows:

$$\begin{aligned}
&Y := [y_1', y_2', ..., y_n']' \text{ where } y_i \in \mathbb{R}^T\ \forall i \in \{1, ..., N\} \\
&X := \begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_N \end{pmatrix} \text{ where } X_i \in \mathbb{R}^{T\times K}\ \forall i \in \{1, ..., N\} \\
&C := (I_N \otimes 1_T) \in \mathbb{R}^{NT\times N} \\
&\eta = [\eta_1, \eta_2, ..., \eta_N]' \\
&V = [v_1', v_2', ..., v_n']' \text{ where } v_i \in \mathbb{R}^T\ \forall i \in \{1, ..., N\} \\
\implies &Y = X\beta + C\eta + V
\end{aligned}$$

We can aim to estimate (β, η) via the following OLS minimization problem

$$\begin{aligned}
&\min_{b,\{\eta_i\}_{i=1}^N} \frac{1}{2}\sum_{i=1}^N (y_i - X_i\beta)'(y_i - X_i b) \\
\iff &\min_{b,\{\eta_i\}_{i=1}^N} \frac{1}{2}\sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it}'b)^2
\end{aligned} \tag{7}$$

Taking the FOC with respect to η_i, I get that

$$\begin{aligned}
0 &= \sum_{t=1}^T (y_{it} - x_{it}'b - \hat{\eta}_i(b)) \\
\implies \hat{\eta}_i(b) &= \frac{1}{T}\sum_{t=1}^T (y_{it} - x_{it}'b) \\
&= \bar{y}_i - \bar{x}_i'b
\end{aligned}$$

[3]Note that strict exogeneity in Assumption (1) is overkill for $\hat{\beta}_N^{FD,OLS}$ to be consistent for β. In fact, a weaker sufficient condition is that $\mathbb{E}[v_{it}|x_{i(t-1)}, x_{it}, x_{i(t+1)}, \eta_i] = 0\ \forall (i,t) \in \{1, ..., N\} \times \{1, ..., T\}$. Thus, while for $T \geq 3$, $\hat{\beta}_N^{FD,OLS}$ is not efficient, it requires less assumptions than $\hat{\beta}_N^{WG,GLS2}$ to be consistent for β

We can substitute this expression back into the minimization problem of equation (7) to get that that problem is equivalent to

$$\begin{aligned}
&\min_b \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'b - \bar{y}_i - \bar{x}_i'b)^2 \\
\iff &\min_b \sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{x}_{it}'b)^2 \\
\implies &\hat{\beta}_N^{WG,GLS2} = \text{argmin}_b \sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{x}_{it}'b)^2
\end{aligned}$$

As a result,

$$\begin{aligned}
\hat{\eta}_i &= \bar{y}_i - \bar{x}_i'\hat{\beta}_N^{WG,GLS2} \\
&= \bar{x}_i'(\beta - \hat{\beta}_N^{WG,GLS2}) + \eta_i + \bar{v}_i \\
&= O_p(\frac{1}{N}) + \eta_i + O_p(\frac{1}{T})
\end{aligned}$$

so that we need T large for $\bar{v}_i \to 0$ so that we can get a could estimate of η_i. Intuitively, we need the time-dimension to be large to learn about the fixed effect, which is time-invariant.

Finally, we can arrive at the estimator $\hat{\beta}_N^{WG,GLS2}$ in one more way. Define the matrix

$$A := (DD')^{-1/2}D \tag{8}$$

so that $Q = A'A$ and $AA' = I_{T-1}$. The within-group estimator can also be seen as the OLS estimate of the regression equation $Ay_i = AX_i\beta + Av_i$ for each $i \in \{1, ..., N\}$ or consolidated $(I_N \otimes A)Y = (I_N \otimes A)X\beta + (I_N \otimes A)V$.

For notational convenience, for any conformable variable r (ie., the first dimension is T), define $r^* := Ar$. We note that for $r \in \mathbb{R}^T$, $r^* \in \mathbb{R}^{T-1}$ and takes the form

$$r_t^* = \sqrt{\frac{T-t}{T-t-1}}\left[r_t - \frac{1}{T-t}(r_{t+1} + ... + r_T)\right] \tag{9}$$

so that this transformation is getting the *forward orthogonal deviations*. Then, for instance $X_i^* = AX_i$, $y_i^* = Ay_i$, and $v_i^* = Av_i$. Also $X^* = (I_N \otimes A)X$, $Y^* = (I_N \otimes A)Y$, and $V^* = (I_N \otimes A)V$. That means I can write the sample regression equations $y_i^* = x_i^*\beta + v_i^*$ for each $i \in \{1, ..., N\}$ and consolidated $Y^* = X^*\beta + V^*$.

By Assumption (1), $\mathbb{E}[V^*|X^*] = 0$ so that the estimator that results from OLS, which is equivalent to $\hat{\beta}_N^{WG,GLS2}$ is consistent for β. That estimator is

$$\hat{\beta}_N^{WG,GLS2} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^* \tag{10}$$

Note that

$$\begin{aligned}\operatorname{Var}(v_i^*|X_i,\eta_i) &= \operatorname{Var}(Av_i|X_i,\eta_i)\\ &= A\operatorname{Var}(v_i|X_i,\eta_i)A'\\ &= A\sigma^2 I_T A'\\ &= \sigma^2 I_{T-1}\end{aligned} \tag{11}$$

which preserves Assumption (2) so that this OLS estimator is also efficient.

1.3.3 Equivalence between First Differences and Within Group when $T = 2$

Recall that I distinguish the first differences estimator when $T = 2$ from when $T \geq 3$. That is because when $T = 2$, the first differences estimator is in fact equivalent to the within-group estimator and therefore is efficient.

Recall the within-group estimator in equation (5). To see the equivalence when $T = 2$, it's useful to simplify $(DD')^{-1} = ([-1,1][-1,1]')^{-1} = \frac{1}{2}$. Thus,

$$\begin{aligned}\hat{\beta}_N^{WG,GLS2} &= \left(\sum_{i=1}^N (X_i'D')(DD')^{-1}(DX_i)\right)^{-1}\left(\sum_{i=1}^N (X_i'D')(DD')^{-1}Dy_i\right)\\ &= \left(\sum_{i=1}^N (X_i'D')(\frac{1}{2})(DX_i)\right)^{-1}\left(\sum_{i=1}^N (X_i'D')(\frac{1}{2})Dy_i\right)\\ &= \left(\sum_{i=1}^N (X_i'D')(DX_i)\right)^{-1}\left(\sum_{i=1}^N (X_i'D')Dy_i\right)\end{aligned}$$

which precisely matches the first differences estimator in equation (4).

1.3.4 Inference and Cluster Robust Standard Errors

Recall from equation (10) that $\hat{\beta}_N^{WG,GLS2} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$.

Under Assumption (2), assuming knowledge of σ^2, we can precisely calculate the finite sample variance of $\hat{\beta}_N^{WG,GLS2}$ as

$$\begin{aligned}\operatorname{Var}(\hat{\beta}_N^{WG,GLS2}|X) &= \operatorname{Var}((X^{*\prime}X)^{-1}X^*Y^*|X)\\ &= (X^{*\prime}X^*)^{-1}X^{*\prime}\operatorname{Var}(Y^*|X)X^*(X^{*\prime}X^*)^{-1}\\ &= (X^{*\prime}X^*)^{-1}X^{*\prime}\operatorname{Var}(V^*|X)X^*(X^{*\prime}X^*)^{-1}\\ &= (X^{*\prime}X^*)^{-1}X^{*\prime}\sigma^2(I_{T-1}\otimes I_N)X^*(X^{*\prime}X^*)^{-1}\\ &= \sigma^2(X^{*\prime}X^*)^{-1}X^{*\prime}(I_{N(T-1)})X^*(X^{*\prime}X^*)^{-1}\\ &= \sigma^2(X^{*\prime}X^*)^{-1}\end{aligned}$$

If we don't have knowledge of σ^2, we can consistently and unbiasedly estimate it as

$$\hat{\sigma}_N^2 := \frac{1}{NT-N-k}(Y^* - X^*\hat{\beta}_N^{WG,GLS2})'(Y^* - X^*\hat{\beta}_N^{WG,GLS2}) \tag{12}$$

However, now suppose that Assumption (2), doesn't hold. In other words, the fourth step of the above derivation isn't kosher. How can we consistently estimate the finite sample variance $\text{Var}(\hat{\beta}_N^{WG,GLS2}|X)$[4]?

$$\begin{aligned}
\hat{\beta}_N^{WG,GLS2} &= (X^{*\prime}X)^{-1}X^{*\prime}Y^* \\
&= (X^{*\prime}X)^{-1}X^{*\prime}(X^*\beta + V^*) \\
&= \beta + (X^{*\prime}X)^{-1}X^{*\prime}V^* \\
\implies (\hat{\beta}_N^{WG,GLS2} - \beta) &= (X^{*\prime}X)^{-1}X^{*\prime}V^* \\
\implies \sqrt{N}(\hat{\beta}_N^{WG,GLS2} - \beta) &= \left(\frac{1}{N}\sum_{i=1}^N X_i^{*\prime}X_i^*\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N X_i^{*\prime}v_i^*\right)
\end{aligned}$$

Under Assumption (1) and finite moment fourth moments, we have that

$$\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N X_i^{*\prime}v_i^*\right) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[X_i^{*\prime}v_i^*(X_i^{*\prime}v_i^*)'])$$

Similarly under finite second moments, and the continuous mapping theorem, we have that

$$\left(\frac{1}{N}\sum_{i=1}^N X_i^{*\prime}X_i^*\right)^{-1} \xrightarrow{\mathbb{P}} \mathbb{E}[X_i^{*\prime}X_i^*]^{-1}$$

Thus, by the delta method, we have that

$$\sqrt{N}(\hat{\beta}_N^{WG,GLS2} - \beta) \xrightarrow{d} \mathcal{N}(0, V)$$

where $V = \mathbb{E}[X_i^{*\prime}X_i^*]^{-1}\mathbb{E}[X_i^{*\prime}v_i^*(X_i^{*\prime}v_i^*)']\mathbb{E}[X_i^{*\prime}X_i^*]^{-1}$. We can in fact consistently estimate all components of V.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N X_i^{*\prime}X_i^* &\xrightarrow{\mathbb{P}} \mathbb{E}[X_i^{*\prime}X_i^*] \\
\frac{1}{N}\sum_{i=1}^N X_i^{*\prime}v_i^*v_i^{*\prime}X_i^* &\xrightarrow{\mathbb{P}} \mathbb{E}[X_i^{*\prime}v_i^*(X_i^{*\prime}v_i^*)']
\end{aligned}$$

Thus, a consistent estimate[5] for the finite sample variance of $\hat{\beta}_N^{WG,GLS2}$ is the *cluster robust* one

$$\begin{aligned}
\hat{V}_N^{CR} &:= \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^N X_i^{*\prime}X_i^*\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N X_i^{*\prime}v_i^*v_i^{*\prime}X_i^*\right)\left(\frac{1}{N}\sum_{i=1}^N X_i^{*\prime}X_i^*\right)^{-1} \\
&= \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T x_{it}^*x_{it}^{*\prime}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T x_{it}^*v_{is}^*v_{it}^*x_{is}^{*\prime}\right)\left(\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T x_{it}^*x_{it}^{*\prime}\right)^{-1}
\end{aligned} \tag{13}$$

[4]Note that calling this estimator GLS under this residual variance assumption is a mistake. We're specifically assuming a different variance structure for the residuals than the one in Assumption (2) that makes this estimator the GLS estimator, which is why the estimator has the superscript '2'.

[5]In the sense that if we estimate $\hat{\text{Var}}_N(\hat{\beta}_N^{WG,GLS2}|X) := \hat{V}_N^{CR}$, then $\lim_{N\to\infty}\text{Var}(\sqrt{N}\hat{\beta}_N^{WG,GLS2}|X) = V$.

Assumption 3 (Unconditional Autocorrelation and Time Series Heteroskedasticity)**.** $\text{Var}(v_i^*|X_i, \eta_i) = \Omega$

Note that Assumption (3) is equivalent to ruling out conditional heteroskedasticity but allows for unconditional autocorrelation and time series heteroskedasticity in the original equation errors v_{it}. It assumes that individual time series are independently and identically distributed with the same variance. Note that this assumption begs the feasible GLS estimator

$$\hat{\beta}_N^{WG,FGLS3} := \left(\sum_{i=1}^{N} X_i^{*\prime}\hat{\Omega}_N^{-1}X_i^*\right)^{-1}(\sum_{i=1}^{N} X_i^{*\prime}\hat{\Omega}_N^{-1}y_i)$$

where the '3' in the definition of $\hat{\beta}_N^{WG,FGLS3}$ refers to the fact that the estimator is the GLS estimator under Assumption (3) and $\hat{\Omega}_N = \frac{1}{N}\sum_{i=1}^{N} v_i^* v_i^{*\prime}$.

Note that the cluster robust variance estimator in (13) holds for balanced panels meaning that the number of observations for each individual i is the same. A more general robust variance estimator for unbalanced panels (ie., individual i can have T_i time series observations) is

$$\hat{V}_N^{CR,UB} := \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} x_{it}^* x_{it}^{*\prime}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{s=1}^{T_i} x_{it}^* v_{is}^* v_{it}^* x_{is}^{*\prime}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} x_{it}^* x_{it}^{*\prime}\right)^{-1}$$

Assumption 4 (Conditional Heteroskedasticity of Unknown Form)**.** $\text{Var}(v_i^*|X_i, \eta_i) = \Omega(X_i, \eta_i)$

There's no clear feasible GLS estimator if we assume the residual variance structure of Assumption (4).

1.4 Likelihood Approaches

The within-group estimator can be viewed as the Gaussian maximum likelihood estimator under three different approaches– joint, condition, and marginal– relative to the individual fixed effects. I will discuss the first two in this section.

1.4.1 Joint Likelihood

Under Assumption (2) with the imposition that the errors come from a zero mean multivariate normal distribution, we have that

$$y_i|X_i, \eta_i \sim \mathcal{N}(X_i\beta + \eta_i 1_T, \sigma^2 I_T) \tag{14}$$

The log conditional density of y_i given x_i and η_i takes the form

$$\log(f_{Y|X,\eta}(y_i|X_i, \eta_i)) \propto -\frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - X_i\beta - \eta_i 1_T)'(y_i - X_i\beta - \eta_i 1_T) \tag{15}$$

So that the conditional log-likelihood of a cross-sectional sample of independent and identically distributed observations is

$$l_J(Y|X;\beta,\sigma^2,\eta) = \sum_{i=1}^{N} \log(f(y_i|X_i,\eta_i)) \tag{16}$$

In light of the standard equivalence between least squares and maximum likelihood estimation in the case of iid gaussian errors, the joint maximization of equation (16) with respect to β, η, σ^2 will give the same estimator $\hat{\beta}_N^{MLE,J}$, for β, as the within-groups estimator $\hat{\beta}_N^{WG,GLS2}$. As a remark, note that the MLE estimator of σ^2 is

$$\hat{\sigma}^2_{N,MLE,J} = \frac{1}{NT}\sum_{i=1}^{N}(y_i - X_i\hat{\beta} - \hat{\eta}_i 1_T)'(y_i - X_i\hat{\beta} - \hat{\eta}_i 1_T)$$

Since $\mathbb{E}[\sum_{i=1}^{N}(y_i - X_i\hat{\beta} - \hat{\eta}_i 1_T)'(y_i - X_i\hat{\beta} - \hat{\eta}_i 1_T)] = (NT - N - k)\sigma^2$, we have that $\hat{\sigma}^2_{N,MLE,J}$ is inconsistent for σ^2.

$$\lim_{N\to\infty} \hat{\sigma}^2_{N,MLE,J} = \frac{T-1}{T}\sigma^2$$

This inconsistency is a consequence of the *incidental parameters problem* where the maximum likelihood estimator need not be consistent when the likelihood of the sample depends on a set of parameters that increases with the sample size. The incidental parameters in this case are the $\eta_1, ..., \eta_N$.

1.4.2 Conditional Likelihood

In the linear static model, $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for η_i in the sense that the density of y_i given x_i, η_i, and $\bar{y}_i$ does not depend on η_i. In math

$$f_{Y|X,\eta,\bar{Y}}(y_i|X_i,\eta_i,\bar{y}_i) = f_{Y|X,\bar{Y}}(y_i|X_i,\bar{y}_i)$$

To see this, first note that we can factor the conditional distribution $f_{Y|X,\eta,\bar{Y}}(y_i|x_i,\eta_i,\bar{y}_i)$ in the following useful manner

$$f_{Y|X,\eta,\bar{Y}}(y_i|X_i,\eta_i,\bar{y}_i) = \frac{f_{Y|X,\eta}(y_i|X_i,\eta_i)}{f_{\bar{Y}|X,\eta}(\bar{y}_i|X_i,\eta_i)} \tag{17}$$

Also, under equation (14), we have the following distribution for $\bar{y}_i$.

$$\begin{aligned} \bar{y}_i|X_i,\eta_i &\sim \mathcal{N}(\bar{X}_i'\beta + \eta_i, \frac{\sigma^2}{T}) \\ \implies \log(f(\bar{y}_i|X_i,\eta_i)) &\propto -\frac{1}{2}\log(\sigma^2) - \frac{T}{2\sigma^2}(\bar{y}_i - \bar{X}_i{}'\beta - \eta_i)^2 \end{aligned} \tag{18}$$

In light of equation (17), subtracting equation (18) from equation (15), I get that

$$\log(f_{Y|X,\eta,\bar{Y}}(y_i|X_i,\eta_i,\bar{y}_i)) = -\frac{T-1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}((y_{it}-x_{it}'\beta-\eta_i)-(\bar{y}_i-\bar{X}_i'\beta-\eta_i))^2$$
$$= -\frac{T-1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}((y_{it}-x_{it}'\beta)-(\bar{y}_i-\bar{X}_i'\beta))^2$$

which is not a function of η_i. The conditional log-likelihood of a sample (Y, X) can be written as

$$l_C(Y|X,\bar{Y};\beta,\sigma^2) = \sum_{i=1}^{N}\log(f_{Y|X,\bar{Y}}(y_i|X_i,\bar{y}_i)) \tag{19}$$

The joint maximization of equation (19) with respect to β, σ^2 will give the same estimator $\hat{\beta}_N^{MLE,C}$, for β, as the within-groups estimator $\hat{\beta}_N^{WG,GLS2}$. As for the estimator $\hat{\sigma}^2_{N,MLE,C}$ for σ^2, note that

$$\hat{\sigma}^2_{N,MLE,C} = \frac{1}{N(T-1)}\sum_{i=1}^{N}(y_i - X_i\beta - (\bar{y}_i - \bar{X}_i'\beta))'(y_i - X_i\beta - (\bar{y}_i - \bar{X}_i'\beta))$$

Note that $\hat{\sigma}^2_{N,MLE,C}$ is consistent for σ^2 but not unbiased for σ^2 with finite T.

1.5 Error Components

A major motivation for using panel data is the possibility of separating out permanent from transitory components of variation. As a notational remark, if I say that $X \sim F(a,b)$, I mean that X is a random variable distributed according to F which has mean a and variance b.

1.5.1 A Variance Decomposition

Consider a variance-components model of the form

$$y_{it} = \mu + \eta_i + v_{it}$$

Assumption 5 (Simple Variance Components Model)**.** $\eta_i \overset{iid}{\sim} F(0,\sigma_\eta^2), v_{it} \overset{iid}{\sim} G(0,\sigma_v^2), \eta_i \perp\!\!\!\perp v_{it}\ \forall(t,i)$.

Under Assumption (5), we note that

$$\mathrm{Var}(y_{it}) = \sigma_\eta^2 + \sigma_v^2$$
$$\mathrm{Var}(y_{it}|\eta_i) = \sigma_v^2$$
$$\mathrm{Cov}(y_{it}, y_{i(t-1)}) = \sigma_\eta^2$$

When estimating this model from data, we wish to estimate $(\mu, \eta_i, \sigma_\eta^2, \sigma_v^2)$. Some natural estimates are

$$\begin{aligned}
\hat{\mu} &= \bar{y} \\
\hat{\eta}_i &= \bar{y}_i - \bar{y} \\
\hat{\sigma}_v^2 &= \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2 \\
\hat{\sigma}_\eta^2 &= \frac{1}{N} \sum_{i=1}^{N} (\bar{y}_i - \bar{y})^2 - \frac{\hat{\sigma}_v^2}{T}
\end{aligned}$$

where $\bar{x}_i = \frac{1}{T} \sum_{t=1}^{T} x_{it}$ and $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} \bar{x}_i$, for any $x \in \{y, \eta, v\}$. We first note that

$$\begin{aligned}
\hat{\mu} &= \bar{y} \\
&= \mu + \frac{1}{N} \sum_{i=1}^{N} \eta_i + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} \\
&\to \mu \text{ with } N \to \infty \text{ and fixed } T \text{ under Assumption (5) and invoking the LLN}
\end{aligned}$$

As for $\hat{\eta}_i$,

$$\begin{aligned}
\hat{\eta}_i &= \bar{y}_i - \bar{y} \\
&= (\mu + \eta_i + \bar{v}_i) - (\mu + \bar{\eta} + \bar{v}) \\
&= \eta_i - \bar{\eta} + \bar{v}_i - \bar{v} \\
&\to \eta_i + \bar{v}_i \text{ with } N \to \infty \text{ and fixed } T \text{ under Assumption (5) and invoking the LLN}
\end{aligned}$$

As for $\hat{\sigma}_v^2$,

$$\begin{aligned}
\hat{\sigma}_v^2 &= \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2 \\
&= \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} (\mu + \eta_i + v_{it} - \mu - \eta_i + \bar{v}_i)^2 \\
&= \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} (v_{it} - \bar{v}_i)^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{T-1} \sum_{t=1}^{T} (v_{it} - \bar{v}_i)^2 \right) \\
&\to \sigma_v^2 \text{ with } N \to \infty \text{ and fixed } T \text{ under Assumption (5) and invoking the LLN}
\end{aligned}$$

That is because in the second to last line, for each $i \in \{1, ..., N\}$, each parenthesized term unbiasedly estimates σ_v^2. As for $\hat{\sigma}_\eta^2$, we first see that

$$\begin{aligned}\text{Var}(\bar{y}_i) &= \text{Var}(\eta_i + \bar{v}_i) \\ &= \text{Var}(\eta_i) + \text{Var}(\bar{v}_i), \text{ using Assumption (5)} \\ &= \sigma_\eta^2 + \frac{\sigma_v^2}{T} \\ &=: \bar{\sigma}^2\end{aligned}$$

Then, we have that

$$\begin{aligned}\hat{\sigma}_\eta^2 &= \frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i - \bar{y})^2 - \frac{\hat{\sigma}_v^2}{T} \\ &\to (\sigma_\eta^2 + \frac{\sigma_v^2}{T}) - \frac{\sigma_v^2}{T} \\ &= \sigma_\eta^2 \text{ with } N \to \infty \text{ and fixed } T \text{ under Assumption (5) and invoking the LLN}\end{aligned}$$

We note that even though $\hat{\sigma}_\eta^2$ is consistent for σ_η^2, for finite N, $\hat{\sigma}_\eta^2$ can be negative, which is undesirable for a variance estimator.

Next, for any $i \in \{1, ..., N\}$, define $y_i := [y_{i1}, ..., y_{iT}]'$. We see that for some distribution $F_1(\cdot, \cdot)$, $y_i \overset{iid}{\sim} F_1(\mu 1_T, \Omega)$ where $\Omega = \sigma_v^2 I_T + \sigma_\eta^2 1_T 1_T'$.

I make a few summary remarks on the above. With N large and T small, we can obtain accurate estimates of σ_η^2 and σ_v^2 but not of the individual realizations η_i. For N small and T large, we can obtain accurate estimates of σ_v^2 but not of σ_η^2. In this second case, $\hat{\mu} \approx \mu + \bar{\eta}$ and $\hat{\eta}_i \approx \eta_i - \bar{\eta}$ so that we our estimate of $\mu + \eta_i$ is consistent but we can't consistently estimate either term.

1.5.2 Likelihood Approach

For the $A \in \mathbb{R}^{(T-1)\times T}$ matrix defined in equation (8), define,

$$H := \begin{pmatrix} T^{-1}1_T' \\ A \end{pmatrix} \in \mathbb{R}^{T\times T} \tag{20}$$

We see that for some distribution F_2,

$$Hy_i = [\bar{y}_i, y_i^*]' \overset{iid}{\sim} F_2\left(\begin{pmatrix}\mu \\ 0\end{pmatrix}, \begin{pmatrix}\bar{\sigma}^2 & 0 \\ 0 & \sigma_v^2 I_{T-1}\end{pmatrix}\right)$$

The off diagonal entries are 0 in the covariance matrix because

$$\begin{aligned}
\text{Cov}(y_i^*, \bar{y}_i) &= \text{Cov}(Ay_i, \frac{1}{T}1_T' y_i) \\
&= A\,\text{Cov}(y_i, y_i)\frac{1}{T}1_T \\
&= A\,\text{Cov}(y_i, y_i)\frac{1}{T}1_T \\
&= A(\sigma_v^2 I_T + \sigma_\eta^2 1_T 1_T')\frac{1}{T}1_t \\
&= \frac{\sigma_v^2}{T}A1_T + \frac{\sigma_\eta^2}{T}(A1_t)1_T' 1_T \\
&= 0, \text{ since } A1_T = 0 \text{ by equation (9)}
\end{aligned} \tag{21}$$

Under the assumption that F_2 is multivariate normal, which arises if we take F_1 (defined just above) to be multivariate normal, then we can decompose the normal density as

$$\log(f(y_i)) = \log(f(\bar{y}_i)) + \log(f(y_i^*)) + \log(|det(H)|)$$

Then, the log-likelihood of $D := \{(y_1, ..., y_N)\}$ is given by

$$l(\mu, \bar{\sigma}^2, \sigma_v^2) \propto l_B(\mu, \bar{\sigma}^2) + l_W(\sigma_v^2) \tag{22}$$

where

$$\begin{aligned}
l_B(\mu, \bar{\sigma}^2) &\propto -\frac{N}{2}\log(\bar{\sigma}^2) - \frac{1}{2\bar{\sigma}^2}\sum_{i=1}^{N}(\bar{y}_i - \mu)^2 \\
l_W(\sigma_v^2) &\propto \frac{-N(T-1)}{2}\log(\sigma_v^2) - \frac{1}{2\sigma_v^2}\sum_{i=1}^{N} y_i^{*\prime} y_i^* \\
&= \frac{-N(T-1)}{2}\log(\sigma_v^2) - \frac{1}{2\sigma_v^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i)^2
\end{aligned}$$

If we do maximum likelihood estimation on the likelihood in equation (22), our estimates of $(\mu, \sigma_v^2, \sigma_\eta^2)$ will precisely match those in section 1.5.1.

1.6 Error Components Regression

One is often interested in the analysis of error-components given some conditioning variables that may be time-varying or time-invariant, or both. For instance, we may be interested in teasing out permanent and transitory components of individual earnings by labor market experience and educational categories. This gives rise to a regression version of the previous model where in principle not only μ but also σ_η^2 and σ_v^2 could be functions of x_{it}, f_i, the time-varying and time-invariant regressors. In the standard model, we take μ a linear function of x_{it} and f_i and make it so that the variances do not depend on the parameters.

Error-Components Regression Model

$$y_{it} = x_{it}'\beta + f_i'\gamma + u_{it}$$
$$u_{it} = \eta_i + v_{it}$$
$$u_i|\omega_i \overset{iid}{\sim} F_3(0, \sigma_v^2 I_T + \sigma_\eta^2 1_T 1_T')$$
for $u_i = [u_{i1}, u_{i2}, ..., u_{iT}]'$ and $w_i = [x_{i1}', ..., x_{iT}', f_i']'$

Note that this model is a specialization of the standard fixed effects model except with the additional assumptions that $\mathbb{E}[\eta_i|w_i] = 0$ and $\text{Var}(\eta_i|w_i) = \sigma_\eta^2$.

Given data $D_N := \{(y_{i1}, ..., y_{it}, x_{i1}, x_{iT}, f_i)\}_{i=1}^N$, we note that OLS is an unbiased, consistent but inefficient estimator of $\delta = [\beta', \gamma']'$ where the design matrix for an individual i is $W_i = [X_i, f_i 1_T]$.

$$\hat{\delta}_N^{OLS} = \left(\sum_{i=1}^N W_i' W_i\right)^{-1} \left(\sum_{i=1}^N W_i' y_i\right)$$

The GLS estimator in this case, where $\Omega = \sigma_v^2 I_T + \sigma_\eta^2 1_T 1_T'$, is given by

$$\hat{\delta}_N^{GLS} = \left(\sum_{i=1}^N W_i' \Omega^{-1} W_i\right)^{-1} \left(\sum_{i=1}^N W_i' \Omega^{-1} y_i\right)$$

This estimator is unfeasible so that typically we replace the relevant variance parameters of Ω with feasible estimates

$$\hat{\sigma}_v^2 = \frac{1}{N(T-1) - k} \sum_{i=1}^N \sum_{t=1}^T (\tilde{y}_{it} - \tilde{x}_{it}' \hat{\beta}_N^{WG,FGLS2})^2$$
$$\hat{\sigma}_\eta^2 = \frac{1}{N} \sum_{i=1}^N (\bar{y}_i - \bar{w}_i' \hat{\delta}_N^{BG})^2 - \frac{\hat{\sigma}_v^2}{T}$$

where the estimator $\hat{\sigma}_v^2$ is the same as the one in equation (12). In addition, $\tilde{z}_{it} := z_{it} - \bar{z}_i$ for $z \in \{x, y\}$ and $\hat{\delta}_N^{BG} = \left(\sum_{i=1}^N \bar{w}_i \bar{w}_i'\right)^{-1} \left(\sum_{i=1}^N \bar{w}_i \bar{y}_i\right)$ where BG refers to the 'between-group' estimator.

1.6.1 Likelihood Approach

For the H matrix defined in equation (20), we have that for some distribution F_4,

$$Hy_i|w_i = [\bar{y}_i, y_i^*]'|w_i \overset{iid}{\sim} F_4\left(\begin{pmatrix} \bar{x}_i'\beta + f_i'\gamma \\ X_i^*\beta \end{pmatrix}, \begin{pmatrix} \bar{\sigma}^2 & 0 \\ 0 & \sigma_v^2 I_{T-1} \end{pmatrix}\right)$$

The off diagonal entries are 0 using a similar argument to the one that terminates in equation (21). Under the assumption that F_4 is multivariate normal, which arises if F_3 in the 'Error-Components Regression Model' is multivariate normal, we get that note that the density of $y_i|w_i$ can be decomposed as

$$\log(f(y_i|w_i)) = \log(f(\bar{y}_i|w_i)) + \log(f(y_i^*|w_i)) + \log(|det(H)|)$$

so that the log-likelihood can be decomposed as the sum of the *between* and *with-in* log-likelihoods.

$$l(\beta, \gamma, \sigma_v^2, \bar{\sigma}^2) \propto l_B(\beta, \gamma, \bar{\sigma}^2) + l_W(\beta, \sigma_v^2)$$

where

$$l_B(\beta, \gamma, \bar{\sigma}^2) = \frac{-N}{2}\log(\bar{\sigma}^2) - \frac{1}{2\bar{\sigma}^2}\sum_{i=1}^{N}(\bar{y}_i - \bar{x}_i'\beta - f_i'\gamma)^2$$
$$l_W(\beta, \sigma_v^2) \propto \frac{-N(T-1)}{2}\log(\sigma_v^2) - \frac{1}{2\sigma_v^2}\sum_{i=1}^{N}(y_i^* - X_i^*\beta)'(y_i^* - X_i^*\beta)$$

Separate maximization of l_W and l_B gives rise to the within-group and between group estimators, respectively. The error-components model can be seen as enforcing the restriction that the parameter β is agreed upon between the two estimators.

2 Dynamic Panel Models

The methods discussed here are motivated by time series properties of short panels. For instance, we may be interested in separating out permanent from transitory components of variation as in earnings mobility studies. We may also be interested in a predictive future distribution given a past distribution. For instance, we may be interested in the predictive distribution of future earnings given past earnings.

2.1 Distinguishing Unobserved Heterogeneity from Genuine Dynamics

Consider two simple error component models,

$$y_{it} = \eta_i + v_{it},\ v_{it}|\eta_i \overset{iid}{\sim} (0, \sigma_v^2),\ \eta_i \overset{iid}{\sim} (0, \sigma_\eta^2)$$
$$y_{it} = \eta + v_{it},\ v_{it} = \alpha v_{i(t-1)} + e_{it},\ e_{it} \overset{iid}{\sim} F(0, \sigma_e^2),\ \alpha \in (0, 1)$$

The first model expresses unobserved heterogeneity whereas the second describes dynamics of an outcome. As an example, suppose that we're an insurance company and we have two periods of data and we're trying to set the price of insurance for various individuals. We note that many individuals who have an accident in the first period also have an accident in the second period. We cannot be certain if that's because individuals that got into accidents in both periods are intrinsically reckless or if getting into an accident in the first period changes the state of individuals in such a way that they're more likely to get into an accident in the second period too.

Mathematically, suppose that we observe data for $T = 2$ periods. In the unobserved heterogeneity model, we note that $\mathrm{Cov}(y_{i2}, y_{i1}) = \sigma_\eta^2$ and $\mathrm{Var}(y_{i1}) = \sigma_\eta^2 + \sigma_v^2$.

For the dynamics model, additionally suppose that $v_{i1} \overset{iid}{\sim} (0, \frac{\sigma_e^2}{1-\alpha^2})$ so that it is as if v_{i1} is selected from the long-term distribution of v_{it}. Then, here we have that $\mathrm{Cov}(y_{i1}, y_{i2}) = \alpha\frac{\sigma_e^2}{1-\alpha^2}$ and $\mathrm{Var}(y_{i1}) = \frac{\sigma_e^2}{1-\alpha^2}$.

In both models, we have two free parameters and two cross-sectional averages so that each model is just identified and we cannot distinguish between the two. Note that the dynamics model assumes that $\alpha \in (0, 1)$. If it were the case that $\alpha \in (-1, 0)$, then the dynamics would work in the opposite direction as the unobserved heterogeneity so that one could distinguish between the two.

If it were the case that we had $T = 3$ samples, then in the unobserved heterogeneity case, we would have $\text{Cov}(y_{i1}, y_{i3}) = \sigma_\eta^2$ and in the dynamics case $\text{Cov}(y_{i1}, y_{i3}) = \alpha^2 \frac{\sigma_e^2}{1-\alpha^2}$. With the over-identifying additional restriction, we could test between both models.

Such an issue does generalize. We cannot distinguish between an $MA(p-1)$ process with unobserved heterogeneity from a $MA(p)$ process with no unobserved heterogeneity.

2.2 Time Effects

Often, a time-series analysis of an individual time series will only be meaningful after conditioning on common features. For instance, in consumption models, properties of consumption and income were investigated after conditioning on trends and demographic characteristics of the households. In other instances, it may be important to remove business cycle or seasonal effects in order to avoid confusion between aggregate and individual dynamics. One might consider decomposing the target into an aggregate component and an individual specific component

$$y_{it} = y_t^a + y_{it}^I$$

and specify time series models for y_t^a and y_{it}^I. For instance, if $y_t^a \overset{iid}{\sim} (0, \sigma_a^2)$ follows the basic zero mean error component model $y_{it}^I = \eta_i + v_{it}$. We get that

$$\text{Var}(y) = \sigma_v^2 I_{NT} + \sigma_\eta^2 (I_N \otimes 1_T 1_T') + \sigma_a^2 (1_N 1_N' \otimes I_T)$$

where $y = [y_1', ..., y_N']'$. Stochastic modeling of y_t^a and η_i require large T and large N. In panels, with small N and large T, individual effects can be treated as parameters where as in panels with large N and small T, we can specify a set of T time dummies to estimate. There are many other models that we can consider, for instance individual specific trends in η_i so that $\eta_{it} = \eta_{0i} + \eta_{1i} t$, which will suggest a different covariance matrix for y.

2.3 Estimating the Covariance Structure

The models above indicate a structure on the data covariance matrix. We now abstract from mean components for simplicity and assume that for $y_i \in \mathbb{R}^p$, and some vector of deep parameters $\theta \in \mathbb{R}^K$,

$$\mathbb{E}_{\mathcal{P}}[y_i y_i'] = \Omega(\theta)$$

If y_i is a scalar time series, then $p = T$, else if $y_{it} \in \mathbb{R}^m$, then $p = mT$. Denoting $vecL_R(\cdot)$ as the function that vectorizes the lower-triangular block of its input matrix in row-major order. In turn, we can write the population moments

$$\begin{aligned} vecL_R(\mathbb{E}_{\mathcal{P}}[y_i y_i' - \Omega(\theta)]) &= 0_{\frac{P(P-1)}{2}} \\ \mathbb{E}_{\mathcal{P}}[s_i - \omega(\theta)] &= \end{aligned}$$

where $s_i = vecL_R(y_i y_i')$ and $\omega(\theta) = vecL_R(\Omega(\theta))$. If $\frac{P(P-1)}{2} \geq dim(\theta)$, then we can efficiently estimate θ using the empirical analogue of the population moments with GMM

$$\hat{\theta}_N = \text{argmin}_{c \in \mathbb{R}^K} - \frac{1}{2} (\bar{s}_N - \omega(c))' W (\bar{s}_N - \omega(c))$$

for some positive definite weighting matrix W. We define $\bar{s}_N = \frac{1}{N} \sum_{i=1}^N s_i$. To be efficient, we wish to pick $W = (\text{Cov}(s_i))^{-1}$ which is unknown but can be consistently estimated using $\hat{V} := \hat{\text{Cov}}(s_i) = \frac{1}{N}(\sum_{i=1}^N s_i s_i') - \bar{s}_N \bar{s}_N'$. Defining $H(\theta) := \frac{\partial \omega(\theta)}{\partial \theta'}$, we have that the first order conditions of the minimization problem are

$$0_K = -H(\hat{\theta}_N)' \hat{V}^{-1} (\bar{s}_N - \omega(\hat{\theta}_N))$$

Using standard results from GMM theory, we have that

$$\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} \mathcal{N}(0, U)$$

where $U := (H(\theta)' V^{-1} H(\theta))^{-1}$.

2.4 Autoregressive Models With Individual Effects

Suppose that we have a sample $\{y_{i0}, ..., y_{iT}; \eta_i\}_{i=1}^N$ where

$$y_{it} = \alpha y_{i(t-1)} + \eta_i + v_{it}, \ t \in \{1, ..., T\}, \ |\alpha| < 1 \tag{23}$$

Assumption 6 (New Innovations)**.** $\mathbb{E}[v_{it} | y_i^{t-1}, \eta_i] = 0$ where $y_i^{t-1} = [y_{i0}, ..., y_{i(t-1)}]'$.

We observe y_i^T but not the individual intercept η_i, which can be regarded as a missing time-invariant variable with $\mathbb{E}[\eta_i] = \eta$ and $\text{Var}(\eta_i) = \sigma_\eta^2$.

Note that assumption (6) implies all of the following

$$\begin{aligned}
&\mathbb{E}[y_{it} | y_i^{t-1}, \eta_i] = \alpha y_{i(t-1)} + \eta_i \\
&\mathbb{E}[v_{it} v_{is}] = 0 \ \forall t \neq s \in \{1, ..., T\} \\
&\mathbb{E}[v_{it}] = 0 \ \forall t \in \{1, ..., T\}
\end{aligned}$$

Assumption 7 (Conditional Time-Series Homoskedasticity)**.** $\mathbb{E}[v_{it}^2 | y_i^{t-1}, \eta_i] = \sigma_t^2$.

Assumption 8 (Time-Series Homoskedasticity)**.** $\mathbb{E}[v_{it}^2] = \sigma^2$.

Note that assumption (8) is compatible with something of the form $\mathbb{E}[v_{it}^2 | \eta_i] = \sigma_i^2$. Assuming that $|\alpha| < 1$ guarrantees that the process is stable but not necessarily stationary since stationarity requires that the process started in the distant past, or equivalently that the distribution of initial states coincides with the steady state distribution of the process. Solving equation (23) recursively, I obtain that

$$y_{it} = \left(\sum_{s=0}^{t-1} \alpha^s\right) \eta_i + \alpha^t y_{i0} + \sum_{s=0}^{t-1} \alpha^s v_{i(t-s)}$$
$$\implies \mathbb{E}[y_{it}|\eta_i] = \left(\sum_{s=0}^{t-1} \alpha^s\right) \eta_i + \alpha^t \mathbb{E}[y_{i0}|\eta_i], \text{ under assumption (6)}$$

With $|\alpha| < 1$, by assumption, $\mathbb{E}[y_{it}|\eta_i] \to \mu_i := \frac{\eta_i}{1-\alpha}$ as $t \to \infty$. We refer to μ_i as the *steady state mean* for individual i so that stationarity in mean requires $\mathbb{E}[y_{i0}|\eta_i] = \frac{\eta_i}{1-\alpha}$. Similarly, under assumption (6) and assumption (8), we have that

$$\text{Cov}(y_{it}, y_{i(t-j)}) = \alpha^{2t-j} \text{Var}(y_{i0}|\eta_i) + \alpha^j (\sum_{s=0}^{t-j-1} \alpha^{2s})\sigma^2$$

which for $|\alpha| < 1$ tends to the steady-state j-th autocovariance for individual i given by $\frac{\alpha^j \sigma^2}{1-\alpha^2}$. Thus, under assumption (8), *covariance stationarity* requires that $\text{Var}(y_{i0}|\eta_i) = \frac{\sigma^2}{1-\alpha^2}$, in which case all autocovariances are time-invariant and coincide with the steady state autocovariances.

Assumption 9 (Mean and Covariance Stationarity)**.** $\mathbb{E}[y_{i0}|\eta_i] = \frac{\eta_i}{1-\alpha}$ and $\text{Var}(y_{i0}|\eta_i) = \frac{\sigma^2}{1-\alpha^2}$.

2.4.1 Nickell Bias

Let $y_i = [y_{i1}, ..., y_{iT}]'$ and $y_{i(-1)} = [y_{i0}, ..., y_{i(T-1)}]'$. Then, I can write the within-group estimator of α, under the estimating assumption $\mathbb{E}[(v_i - \bar{v}_i)(y_{i(-1)} - \bar{y}_{i(-1)})] = 0$:

$$\begin{aligned}
\hat{\alpha}_N^{WG} &= \frac{\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)(y_{i(t-1)} - \bar{y}_{i(-1)})}{\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)^2} \\
&= \alpha + \frac{\sum_{i=1}^N \sum_{t=1}^T (v_{it} - \bar{v}_i)(y_{i(t-1)} - \bar{y}_{i(-1)})}{\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)^2} \\
&\not\to \alpha \text{ as } N \to \infty
\end{aligned}$$

The reality is that the estimating assumption does not hold as we do not have strict exogeneity. For intuition, when $\alpha \in (0, 1)$, suppose that $y_{i(t-1)}$ is below it's long run average $\bar{y}_{i(-1)}$, then, we know that $y_{i(t-1)}$ needs to average out to $\bar{y}_{i(-1)}$ so that we will have some positive shocks for that to happen. Thus, for finite T, v_{it} covaries negatively with $y_{i(t-1)}$ and the numerator is negative so that indeed we have a downwards bias.

A typical attempt at showing that $\hat{\alpha}_N^{WG}$ is unbiased would go as follows

$$\begin{aligned}
\mathbb{E}[\hat{\alpha}_N^{WG}] &= \mathbb{E}\left[\alpha + \frac{\sum_{i=1}^N \sum_{t=1}^T (v_{it} - \bar{v}_i)(y_{i(t-1)} - \bar{y}_{i(-1)})}{\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)^2}\right] \\
&= \alpha + \mathbb{E}\left[\frac{\sum_{i=1}^N \sum_{t=1}^T (y_{i(t-1)} - \bar{y}_{i(-1)}) \mathbb{E}[v_{it} - \bar{v}_i | y_{i(-1)}]}{\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)^2}\right]
\end{aligned}$$

One would attempt to argue that $\mathbb{E}[v_{it} - \bar{v}_i | y_{i(-1)}] = 0$ for any (i, t). However that is not true here since given future values of y_{it} we can say a lot about $v_{it} - \bar{v}_i$ (the lack of strict exogeneity here is key). One can then show that,

$$\begin{aligned}\lim_{N\to\infty}\hat{\alpha}_N^{WG}-\alpha &= \frac{\mathbb{E}[\sum_{t=1}^{T}(v_{it}-\bar{v}_i)(y_{i(t-1)}-\bar{y}_{i(-1)})]}{\mathbb{E}[\sum_{t=1}^{T}(y_{it}-\bar{y}_i)^2]}\\ &= -\frac{(1-\alpha^2)h_T(\alpha)}{T-1}\left(1-\frac{2\alpha h_T(\alpha)}{T-1}\right)^{-1}, \text{ under assumptions (6), (8), and (9)}\end{aligned}$$

where $h_T(\alpha) := \frac{1}{1-\alpha}\left[1-\frac{1}{T}\left(\frac{1-\alpha^T}{1-\alpha}\right)\right]$. One notes that the asymptotic bias (asymptotic bias in N) is of order $\frac{1}{T}$ and vanishes as $T\to\infty$ but for small T it can be sizeable.

2.4.2 Using Instruments on First Differences

We note that we can write the model in equation (23) after first differences as

$$\Delta y_{it} = \alpha\Delta y_{i(t-1)} + \Delta v_{it}$$

Running OLS on this regression will again produce an asymptotically biased estimator of α since strict exogeneity doesn't hold. However, we can consider using an instrumental variables approach to get a consistent estimator for α. Consider the candidate instrument, $y_{i(t-2)}$. Note that

$$\begin{aligned}\mathbb{E}[y_{i(t-2)}\Delta v_{it}] &= \mathbb{E}[\mathbb{E}[y_{i(t-2)}(v_{it}-v_{i(t-1)})|y_i^{t-2}]]\\ &= \mathbb{E}[y_{i(t-2)}\underbrace{\mathbb{E}[(v_{it}-v_{i(t-1)})|y_i^{t-2}]}_{0}]\\ &= 0\\ \mathbb{E}[y_{i(t-2)}\Delta y_{i(t-1)}] &= \mathbb{E}[y_{i(t-2)}(y_{i(t-1)}-y_{i(t-2)})]\\ &= \mathbb{E}[y_{i(t-2)}(\alpha y_{i(t-2)}+\eta_i+v_{i(t-1)}-y_{i(t-2)})]\\ &= (\alpha-1)\mathbb{E}[y_{i(t-2)}^2]+\eta_i\mathbb{E}[y_{i(t-2)}]+\underbrace{\mathbb{E}[y_{i(t-2)}v_{i(t-1)}]}_{0}\\ &= (\alpha-1)\mathbb{E}[y_{i(t-2)}^2]+\eta_i\mathbb{E}[y_{i(t-2)}]\\ &\neq 0\end{aligned}$$

unless we have some knife edge case. Thus, indeed, we have that the relevance and inclusion IV restrictions are satisfied and that $\mathbb{E}[y_{i(t-2)}\Delta v_{it}]=0$ is an estimating assumption that will yield a consistent estimate of α on the first differences regression. In fact, we can use any past occurrence $y_{is}\in y_i^{t-2}$ as an instrument. Compactly, I will write that the following $\frac{(T-1)T}{2}$ linear IV moment restrictions hold

$$\mathbb{E}[y_i^{t-2}\Delta v_{it}] = 0$$

Using GMM on these moment conditions, α can be consistently estimated.

3 PEER EFFECTS

The linear model analyzed expresses three potential hypotheses about why individuals in the same group often behave similarly.

(a) *endogenous effects*, the propensity of an individual to behave in some way varies with the behavior of the group

(b) *exogenous effects*, the propensity of an individual to behave in some way varies with the exogenous characteristics of the group; this is also referred to as *contextual effects*

(c) *correlated effects*, individuals in the same group tend to behave similarly because they have similar characteristics or face similar institutional environments.

As an example, consider high-school achievement for teenage youth. There is an endogenous effect if all else equal individual achievement tends to vary with the average achievement of the group. There is an exogenous effect if achievement tends to vary with say, the socio-economic composition of the reference group. There are correlated effects if youths in the same school tend to achieve similarly because they have similar family backgrounds or are taught by the same teachers.

3.1 Model Specification [Manski (1993)]

Let each member of a population be characterized by a value for $(y, x, z, u) \in \mathbb{R}^1 \times \mathbb{R}^J \times \mathbb{R}^K \times \mathbb{R}^1$. In this setting, y is viewed as a scalar outcome (eg., academic achievement), x are attributes characterizing an individual's reference group (eg., a youth's school or ethnic group), and (z, u) are attributes that directly affect y (eg., socio-economic status and ability). We note that (y, u) are the only individual specific variables that we have data on. Meanwhile, (x, z) are attributes of a group. Assume that

$$\begin{aligned} & y = \alpha + \mathbb{E}[y|x]\beta + \mathbb{E}[z|x]'\gamma + z'\eta + u,\ \mathbb{E}[u|x,z] = x'\delta \\ \implies & \mathbb{E}[y|x,z] = \alpha + \mathbb{E}[y|x]\beta + \mathbb{E}[z|x]'\gamma + x'\delta + z'\eta \end{aligned} \tag{24}$$

To clarify notation, we say that $\mathbb{E}[y|x]$ is the mean achievement in the reference group, characterized by x, of the individual. Meanwhile, $\mathbb{E}[z|x]$ is viewed as the mean of the exogenous variables z among the persons in the reference group. If $\beta \neq 0$, then the regression expresses an endogenous effect since a person's achievement varies with the mean achievement of the group. If $\gamma \neq 0$, the model expresses an exogenous effect because y varies with the mean of the exogenous variables in the reference group. If $\delta \neq 0$, the model expresses correlated effects because persons in the reference group tend to behave similarly because they have similar unobserved characteristics or face similar institutional environments. Finally, $\eta \neq 0$, would mean that there's a direct effect of z on y.

We can integrate equation (24) over z to get

$$\begin{aligned} & \mathbb{E}[y|x] = \alpha + \mathbb{E}[y|x]\beta + \mathbb{E}[z|x]'\gamma + x'\delta + \mathbb{E}[z|x]'\eta \\ \implies & \mathbb{E}[y|x] = \frac{\alpha}{1-\beta} + \mathbb{E}[z|x]'\left(\frac{\gamma+\eta}{1-\beta}\right) + x'\left(\frac{\delta}{1-\beta}\right),\ \text{if } \beta \neq 1 \end{aligned} \tag{25}$$

Thus, $\mathbb{E}[y|x]$ is a linear function of $[1, \mathbb{E}[z|x]', x']'$. We can plug equation (25) into equation (24) to obtain the reduced form model

$$\mathbb{E}[y|x,z] = \frac{\alpha}{1-\beta} + \mathbb{E}[z|x]'\left[\frac{\gamma+\beta\eta}{1-\beta}\right] + x'\left(\frac{\delta}{1-\beta}\right) + z'\eta$$

In the linear model of equation (24) with $\beta \neq 1$, the composition parameters $(\frac{\alpha}{1-\beta}, \left[\frac{\gamma+\beta\eta}{1-\beta}\right], \frac{\delta}{1-\beta}, \eta)$, are identified if the regressors $[1, \mathbb{E}[z|x], x, z]$ are linearly independent in the population. In these cases, we cannot tease out which peer effect is present due to the existence of these composite parameters.

3.2 Model Specification [Bramoullé et al. (2009)]

We assume that we have data $D := \{(x_i, y_i, N_i)\}_{i=1}^N$ on N individuals. We let y_i be the outcome of interest, x_i be the socio-economic characteristics of individual i, and $N_i \subset \{1, ..., N\}$ is the reference group for individual i of size n_i. The reference group of individual i is assumed to exclude individual i and contains all individuals that may influence the outcome of individual i. We assume that we have our sample D is iid from a population of networks with a fixed and known structure. The structural model is given by

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 \frac{\sum_{j \in N_i} y_j}{n_i} + \alpha_3 \frac{\sum_{j \in N_i} x_j}{n_i} + \epsilon_i$$

where α_2 captures endogenous effects (we impose $|\alpha_2| < 1$) and α_3 captures endogenous effects. We assume that there are no correlated effects such as by randomization of an individual's reference group. We impose this by assuming strict exogeneity, ie., that $\mathbb{E}[\epsilon_i | x] = 0$ where $x = [x_1, ..., x_N]$.

Define the matrix $G \in \mathbb{R}^{N \times N}$ by $G_{ij} = \mathbb{1}_{j \in N_i} \left(\frac{1}{n_i} \right)$. For our sample, the population equation becomes

$$y = a_0 1_N + \alpha_1 x + \alpha_2 G y + \alpha_3 G x + \epsilon \tag{26}$$

where $\mathbb{E}[\epsilon | x] = 0$ and $\mathbb{E}[\epsilon \epsilon'] = \Sigma$. If the matrix $I_N - \alpha_2 G$ is invertible, we can write the above equation as

$$y = \alpha_0 \left(I_N - \alpha_2 G \right)^{-1} 1_N + \left(I - \alpha_2 G \right)^{-1} G x + \left(I - \alpha_2 G \right)^{-1} \epsilon$$

where the intercept is simply $\frac{\alpha_0}{1-\alpha_2}$ if the individual is not isolated and α_0 otherwise. The covariance matrix of the errors is given by $\Omega := ((I - \alpha_2 G)^{-1}) \Sigma ((I - \alpha_2 G)^{-1})' \implies \Sigma = (I - \alpha_2 G) \Omega (I - \alpha_2 G)'$. The social effects are identified if and only if $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ are uniquely recovered from the restricted reduced form parameters.

Since $(I - \alpha_2 G)^{-1} = \sum_{k=0}^{\infty} (\alpha_2)^k G^k$, with the assumption that the eigenvalues of $\alpha_2 G$ are less than 1, and assuming no isolated individuals and given the strict exogeneity assumption,

$$\begin{aligned} y &= \left(\frac{\alpha_0}{1 - \alpha_2} \right) 1_N + \alpha_1 x + (\alpha_1 \alpha_2 + \alpha_3) \sum_{k=0}^{\infty} (\alpha_2)^k G^{k+1} x + \sum_{k=0}^{\infty} (\alpha_2)^k G^k \epsilon \\ \implies \mathbb{E}[Gy | x] &= \left(\frac{\alpha_0}{1 - \alpha_2} \right) 1_N + \alpha_1 G x + (\alpha_1 \alpha_2 + \alpha_3) \sum_{k=0}^{\infty} (\alpha_2)^k G^{k+2} x \end{aligned} \tag{27}$$

There's a proposition in this paper that under the assumption that $\alpha_1 \alpha_2 + \alpha_3 \neq 0$, social effects are identified if and only if the matrices I_N, G, G^2 are linearly independent. This linear independence is intuitive since we need 4 moments to identify the four parameters in the structural equation (26). Those four moments arise from (1) the population, (2) the individual specific effects (ie., I_N), (3) effects from one degree away (ie., G), and (4) effects from 2 degrees away (ie., G^2). The condition that $\alpha_1 \alpha_2 + \alpha_3 \neq 0$ simply expresses the fact that there in fact exists some kind of direct or indirect social effect that one can identify. Equation (27) makes it clear that any of $(1, x, Gx, G^2 x, ...)$ can be used as valid instruments and therefore can be used to consistently estimate the α's.

4 MIXTURE MODELS

We assume that we observe iid data $\{(X_i, Y_i)\}_{i=1}^N$ where each data point comes from one of $k \in \{1, ..., K\}$ unobserved types. We say that the conditional density of $f_{Y|X}$ is given by

$$f_{y|x}(y_i|x_i) = \sum_{k=1}^{K} \Pr(k|x_i) f^k_{Y|X}(y_i|x_i)$$

where $f^k_{Y|X}$ is the conditional density of Y given X if the data point is known to be of type k. As an example, we can consider

$$\begin{aligned}
&y_{i\alpha} := \text{career choice of individual } i \text{ at age } \alpha \\
&y_i := [y_{i(16)}, y_{i(17)}, y_{i(18)}, ..., y_{i(\bar{a})}] \\
&x_i := \text{education of individual } i \text{ at age 16} \\
&k := \text{endowment heterogeneity (unobserved)}
\end{aligned}$$

Assuming that there's no path-dependence in career choices, and the only persistent determinant of an individual's career choice is an individual's unobserved latent type, then

$$f^k_{Y|X}(y_i|x_i) = \Pi^{\bar{a}}_{a=16} f^k_{Y_\alpha|X}(y_{i\alpha}|x_i)$$

Note that in this example, x_i is time-invariant so that we could not distinguish the effect of an individual or group fixed effect on career choices from x. As a result, we're not permitted to estimate a fixed effects model and settle for this random-effects approach based on latent types.

The log-likelihood of a sample is given by

$$l(y|x;\theta,\lambda) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \Pr(k|x_i;\lambda) f^k_{Y|X}(y_i|x_i,\theta_k) \right)$$

We note that the first order conditions with respect to λ involve all parameters so that it's a highly nonlinear system, which is difficult to solve computationally.

4.1 Expectation Maximization Algorithm

We define a new variable $z_i := k_i$ to be the latent type of individual i, assuming correct specification. In the expectation maximization algorithm, we initialize $\theta^{(0)}$ and $\lambda^{(0)}$ and then define the quantity

$$\begin{aligned}
Q(\theta,\lambda|\theta^{(t)},\lambda^{(t)}) &:= \mathbb{E}\left[\sum_{i=1}^{N} \log\left(f_{Y,Z|X}(y_i,z_i|x_i;\theta,\lambda) \right) |y_i,x_i,\theta^{(t)},\lambda^{(t)} \right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} \log\left(f^k_{Y|X}(y_i|x_i;\theta_k)\Pr(k|x_i;\lambda) \right) |y_i,x_i,\theta^{(t)},\lambda^{(t)} \right] \\
&\approx \sum_{i=1}^{N}\sum_{k=1}^{K} \Pr(z_i=k|y_i,x_i,\theta^{(t)},\lambda^{(t)}) \log\left(f^k_{Y|X}(y_i|x_i;\theta_k)\Pr(k|x_i;\lambda) \right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{f^k_{Y|X}(y_i|x_i;\theta^{(t)}_k)\Pr(k|x_i;\lambda^{(t)})}{\sum_{k'=1}^{K} f^{k'}_{Y|X}(y_i|x_i;\theta^{(t)}_{k'})\Pr(k'|x_i;\lambda^{(t)})} \log\left(f^k_{Y|X}(y_i|x_i;\theta_k)\Pr(k|x_i;\lambda) \right)
\end{aligned}$$

Next, we do the maximization step

$$(\theta^{(t+1)}, \lambda^{(t+1)}) = \text{argmax}_{\theta,\lambda} Q(\theta, \lambda | \theta^{(t)}, \lambda^{(t)})$$

We iterate until $||[\theta^{(t+1)}, \lambda^{(t+1)}]' - [\theta^{(t)}, \lambda^{(t)}]'||_2 < \epsilon$, for some chosen convergence value $\epsilon > 0$.

4.1.1 Justification for the Algorithm

For a given $i \in \{1, ..., N\}$, we have that

$$\log\left(f_{Y|X}(y_i|x_i;\theta,\lambda)\right) = \log\left(f_{Y,Z|X}(y_i, z_i|x_i;\theta,\lambda)\right) - \log\left(f_{Z|Y,X}(z_i|y_i, x_i;\theta,\lambda)\right)$$

As a result, for any value of (θ, λ), we have

$$\begin{aligned}
\sum_{i=1}^{N} \log(f_{Y|X}(y_i|x_i;\theta,\lambda)) &= \sum_{i=1}^{N}\sum_{k=1}^{K} \Pr(z_i = k|y_i, x_i; \theta^{(t)}, \lambda^{(t)}) \left[\log\left(f^k_{Y|X}(y_i|x_i;\theta_k)\Pr(k|x_i;\lambda)\right) - \log\left(f_{Z|Y,X}(z_i|y_i, x_i;\theta,\lambda)\right)\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \Pr(z_i = k|y_i, x_i; \theta^{(t)}, \lambda^{(t)}) \log\left(f^k_{Y|X}(y_i|x_i;\theta_k)\Pr(k|x_i;\lambda)\right) \\
&\quad - \sum_{i=1}^{N}\sum_{k=1}^{K} \Pr(z_i = k|y_i, x_i; \theta^{(t)}, \lambda^{(t)}) \log\left(f_{Z|Y,X}(z_i|y_i, x_i;\theta,\lambda)\right) \\
&= Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) + H(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) \qquad (28)
\end{aligned}$$

where $H(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) := -\sum_{i=1}^{N}\sum_{k=1}^{K} \Pr(z_i = k|y_i, x_i; \theta^{(t)}, \lambda^{(t)}) \log\left(f_{Z|Y,X}(z_i|y_i, x_i;\theta,\lambda)\right)$. Also define $F(\theta, \lambda) := \sum_{i=1}^{N} \log(f_{Y|X}(y_i|x_i;\theta,\lambda))$. As a result, we have that

$$\begin{aligned}
F(\theta,\lambda) - F(\theta^{(t)}, \lambda^{(t)}) &= Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) + H(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) - \left(Q(\theta^{(t)}, \lambda^{(t)}|\theta^{(t)}, \lambda^{(t)}) + H(\theta^{(t)}, \lambda^{(t)}|\theta^{(t)}, \lambda^{(t)})\right) \\
&= \left(Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) - Q(\theta^{(t)}, \lambda^{(t)}|\theta^{(t)}, \lambda^{(t)})\right) + \left(H(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) - H(\theta^{(t)}, \lambda^{(t)}|\theta^{(t)}, \lambda^{(t)})\right)
\end{aligned}$$

Next, we make use of the following inequality: for any given $(p_1, ..., p_K) \in \Delta^K_{++}$ and any $(q_1, ..., q_K) \in \Delta^K_{++}$, we have that

$$-\sum_{l=1}^{K} p_l \log(p_l) \leq -\sum_{l=1}^{K} p_l \log(q_l)$$

To see why this holds, we know that $\log(x) \leq x - 1 \; \forall x > 0$. That implies that

$$
\begin{aligned}
-\sum_{l=1}^{K} p_l \log\left(\frac{q_l}{p_l}\right) &\geq -\sum_{k=1}^{K} p_l \left(\frac{q_l}{p_l} - 1\right) \\
&= \sum_{l=1}^{K} p_l - \sum_{l=1}^{K} q_l \\
&= 0 \\
\implies -\sum_{l=1}^{K} p_l \log(p_l) &\leq -\sum_{l=1}^{K} p_l \log(q_l)
\end{aligned}
$$

Using this observation, we deduce that for any (θ, λ)

$$H(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) \geq H(\theta^{(t)}, \lambda^{(t)}|\theta^{(t)}, \lambda^{(t)})$$

Then, combining this fact with the result of equation (28), I get that

$$F(\theta, \lambda) - F(\theta^{(t)}, \lambda^{(t)}) \geq Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) - Q(\theta^{(t)}, \lambda^{(t)}|\theta^{(t)}, \lambda^{(t)})$$

so that maximizing Q is equivalent to maximizing the log-likelihood F.

4.2 Group Heterogeneity and K-Means Clustering

Assume again that we observe iid data $\{(X_i, Y_i)\}_{i=1}^{N}$ where each data point comes from one of $k \in \{1, ..., K\}$ types. We assume that there's a type-specific effect so that $\alpha_i = \alpha_{k(i)}$ where $k(i)$ is the group to which individual i belongs. We like this idea because it's possible that heterogeneity at the individual specific level is to granular. For instance, in the case of schooling, we can imagine that cohorts of people that received similar kinds of education may have correlated income streams due to some type specific effect. The catch here is that we assume that the types of individuals are unobserved so that we must learn them from data.

There are many clustering methods (eg., hierarchical, centroid-based, distributional, density-based, spectral, ...) used to cluster individuals into groups. We focus here on K-means clustering.

The K-means problem takes as input a number K groups of to which to assign individuals. It searches for centroids $(\tilde{h}(1), ..., \tilde{h}(K))$ where $\tilde{h}(l) \in \mathbb{R}^{\dim(X)+\dim(Y)}$ *centroids* and individual assignments $k : \{1, ..., N\} \mapsto \{1, ..., K\}$ so as to solve

$$(\hat{h}(1), ..., \hat{h}(K), \hat{k}) = \text{argmin}_{(\tilde{h}(1),...,\tilde{h}(K),k)} \sum_{i=1}^{N} ||[X_i', Y_i]' - \tilde{h}(k(i))||_2^2$$

4.2.1 Lloyd's Algorithm

Lloyd's algorithm provides an approach to solve for the centroids and the assignment function by an iterative algorithm between assignment and computation of centroids. We start with an initial set of centroids $\tilde{h}(1)^{(0)}, ..., \tilde{h}(K)^{(0)}$. We then assign each individual according to

$$k(i)^{(s)} = \text{argmin}_{l \in \{1,...,K\}} ||[X_i', Y_i]' - \tilde{h}(k(l))^{(s)}||_2^2$$

Then, we update the centroids by

$$(\tilde{h}(1)^{(s+1)}, ..., \tilde{h}(K)^{(s+1)}) = \text{argmin}_{\{\tilde{h}(l)\}_{l=1}^{K}} \frac{1}{N} \sum_{i=1}^{N} ||[X_i, Y_i]' - \tilde{h}(k(i)^{(s)})||_2^2$$

We terminate when $\sum_{l=1}^{K} ||\tilde{h}(l)^{(s+1)} - \tilde{h}(l)^{(s)}||_2^2 < \epsilon$ for some previously chosen convergence tolerance $\epsilon > 0$.

4.2.2 Statistics: Oracle Property

Consider a model with time-invariant heterogeneity, no covariates, and $K = 2$:

$$y_{it} = \alpha^*_{k(i)} + v_{it}, k(i) \in \{1, 2\}$$

A key quantity is the following mis-classification quantity for an individual i:

$$\Pr(\hat{k}(i) = 2|k(i) = 1) = \Pr((\bar{y}_i - \alpha_2)^2 < (\bar{y}_i - \alpha_1)^2|k(i) = 1)$$

If we further assume that $v_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $\alpha_1 < \alpha_2$, then

$$\begin{aligned}
\Pr(\hat{k}(i) = 2|k(i) = 1) &= \Pr((\bar{y}_i - \alpha_2)^2 < (\bar{y}_i - \alpha_1)^2|k(i) = 1) \\
&= \Pr(\bar{v}_i > \frac{\alpha_1 + \alpha_2}{2} - \alpha_1^*) \\
&= 1 - \Phi\left(\frac{\sqrt{T}}{\sigma}\left(\frac{\alpha_1 + \alpha_2}{2} - \alpha_1^*\right)\right) \\
&= 1 - \Phi\left(\frac{\sqrt{T}}{2\sigma}(\alpha_2^* - \alpha_1^*)\right), \text{ if we know true locations of centroids}
\end{aligned}$$

This mis-classification quantity vanishes as $T \to \infty$.

5 NONLINEAR PANEL DATA FIXED EFFECT MODELS

Suppose that we observe iid samples $\{(x_{i1}, ..., x_{iT}, y_{i1}, ..., y_{iT}, \alpha_i)\}_{i=1}^{N}$ where

$$y_{it}|y_i^{t-1}, x_{it}, \alpha_i \overset{iid}{\sim} f(y_{it}|y_i^{t-1}, x_{it}, \alpha_i^0; \theta^0)$$

and f is some parametric distribution. We have that the log-likelihood of the data is

$$\begin{aligned} l(y|x,\alpha,\theta) &= \sum_{i=1}^{N} \log\left(f(y_i|x_i,\alpha_i;\theta)\right) \\ &\approx \sum_{i=1}^{N}\sum_{t=2}^{T} \log\left(f(y_{it}|y_i^{t-1},x_i,\alpha_i;\theta)\right) \end{aligned}$$

where $\alpha = [\alpha_1, ..., \alpha_N]', y = [y_1, ..., y_N]', x = [x_1', ..., x_N']'$. A maximum likelihood estimate of (θ, α) solves the problem

$$(\hat{\theta}_{N,T}^{MLE}, \hat{\alpha}_{N,T}^{MLE}) := \text{argmax}_{\theta,\alpha} l(y|x,\alpha,\theta)$$

We note that $\hat{\theta}_{N,T}^{MLE}$ can be equivalently computed from the problem

$$\begin{aligned} \hat{\theta}_{N,T}^{MLE} &= \arg\max_{\theta} l(y|x,\hat{\alpha}(\theta);\theta) \\ &\text{where } \hat{\alpha}_i(\theta) = \text{argmax}_{\alpha_i} \log(f(y_i|x_i,\alpha_i;\theta)) \end{aligned}$$

Denoting $\theta_T^{MLE} = \lim_{N\to\infty} \hat{\theta}_{N,T}^{MLE}$, we get that $\theta_T^{MLE} \neq \theta^0$ so that the estimator is inconsistent for the true parameter θ^0. The reason for this inconsistency is that for finite T, we have some finite-sample error in our estimate of the individual fixed effect so that the MLE estimate for θ inconsistently estimate θ^0; the maximand is different than the one below, so that the pseudo-true parameter in each case will also be different. If it were in fact the case that we knew the fixed effect, then our estimate for θ would be consistent for θ^0. In other words, defining

$$\tilde{\theta}_{N,T}^{MLE} := \arg\max_{\theta} l(y|x,\alpha^0;\theta)$$

we have that $\lim_{N\to\infty} \tilde{\theta}_{N,T}^{MLE} = \theta^0$. The reason for the inconsistency is very similar in spirit to the Nickell Bias covered in Section 2.4.1.

5.1 Split-Panel Jackknife Estimator

For a more technical treatment of this estimator, see (Dhaene and Jochmans 2015).

As in the previous section, suppose that we observe iid samples $\{(x_{i1}, ..., x_{iT}, y_{i1}, ..., y_{iT}, \alpha_i)\}_{i=1}^N$ where

$$y_{it}|y_i^{t-1}, x_{it}, \alpha_i \overset{iid}{\sim} f(y_{it}|y_i^{t-1}, x_{it}, \alpha_i^0; \theta^0)$$

Assumption 10 (Stationarity and Exponentially Decreasing Mixing Coefficients)**.** We assume that the processes (y_{it}, x_i) are independent across i and stationary and alpha mixing across t with mixing coefficients $a_i(m)$ that are uniformly exponentially decreasing (ie., $\sup_i |a_i(m)| < Cb^m$ for some finite C and $b \in (0,1)$). The idea is that observations from t that are relatively far apart are pretty much independent.

Define $s_{it}(\theta) := \nabla_\theta \log(f(y_{it}|y_i^{t-1}, x_{it}, \alpha_i(\theta);\theta))$, $H_{it}(\theta) := \nabla_{\theta\theta'} = \log(f(y_{it}|y_i^{t-1}, x_{it}, \alpha_i(\theta);\theta))$, and $\Sigma := -\mathbb{E}[H_{it}(\theta^0)]$.

Assumption 11 (Existence of Pseudo-True Parameter and Asymptotic Covariance)**.** We assume that θ_T and Σ exist and $\sqrt{NT}(\hat{\theta}^{MLE}_{N,T} - \theta^{MLE}_T) = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \Sigma^{-1} s_{it}(\theta^0) + o_p(1)$ as $N, T \to \infty$.

Assumption 12 (Constant Bias Shrinking Coefficient)**.** As $T \to \infty$, $\theta^{MLE}_T - \theta^0 = \frac{B_1}{T} + o_p(\frac{1}{T})$ for some $B_1 \in \mathbb{R}$.

Assumption (11) is a standard assumption about the asymptotics of the MLE estimator. Assumption (12) is a critical assumption (for this analysis) about how the bias shrinks as $T \to \infty$. These assumptions jointly imply that

$$\begin{aligned}\sqrt{NT}(\hat{\theta}^{MLE}_{N,T} - \theta^0) &= NT(\hat{\theta}^{MLE}_{N,T} - \theta^{MLE}_T + \theta^{MLE}_T - \theta^0) \\ &= \frac{1}{\sqrt{NT}} \left(\sum_{i=1}^N \sum_{t=1}^T \Sigma^{-1} s_{it}(\theta^0) \right) + \sqrt{NT} \frac{B_1}{T} + o_p\left(\frac{\sqrt{NT}}{T} \right) + o_p(1) \\ &\xrightarrow{d} \mathcal{N}(\rho B_1, \Sigma^{-1})\end{aligned}$$

Since $\frac{1}{\sqrt{NT}} \left(\sum_{i=1}^N \sum_{t=1}^T \Sigma^{-1} s_{it}(\theta^0) \right) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$, $\sqrt{NT}\frac{B_1}{T} \xrightarrow{\mathbb{P}} \rho B_1$ where $\rho = \lim_{N,T\to\infty} \sqrt{\frac{N}{T}}$, and then an application of the continuous mapping theorem.

Next, define a subpanel S to be a proper subset $S \subset \{1, ..., T\}$ so that the elements of S are consecutive integers and $|S| \geq T_{\text{min}}$ where T_{min} is the smallest panel size for which θ^{MLE}_T exists. We now define the maximum likelihood estimator corresponding to a subpanel S as

$$\hat{\theta}^{MLE}_{N,S} := \text{argmax}_\theta \hat{l}_S(y|x; \hat{\alpha}_S(\theta), \theta)$$

where $\hat{l}_S(y|x; \hat{\alpha}_S(\theta), \theta) = \sum_{i=1}^N \sum_{t\in S} \log(f(y_{it}|y_i^{t-1}, x_{it}; \hat{\alpha}_{iS}(\theta), \theta)$ and $\hat{\alpha}_{iS}(\theta) := \text{argmax}_{\alpha_i} \sum_{t\in S} \log(f(y_{it}|y_i^{t-1}, x_{it}, \alpha, \theta))$.

Since by their very definition, subpanels preserve the dependency structure of the full panel, we have that $\lim_{N\to\infty} \hat{\theta}^{MLE}_{N,S} \to \theta^{MLE}_{|S|}$ and as $|S|, T \to \infty$, we get that

$$\begin{aligned}\theta^{MLE}_T - \theta^0 &= \frac{B_1}{T} + o_p\left(\frac{1}{T}\right) \\ \theta^{MLE}_{|S|} - \theta^0 &= \frac{B_1}{|S|} + o_p\left(\frac{1}{|S|}\right) \\ \implies (\theta^{MLE}_{|S|} - \theta^{MLE}_T) &= \frac{B_1}{|S|} - \frac{B_1}{T} - o_p\left(\frac{1}{T}\right) + o_p\left(\frac{1}{|S|}\right) \\ \implies \frac{|S|}{T-|S|}(\theta^{MLE}_{|S|} - \theta^{MLE}_T) &= \frac{B_1}{T} - o_p\left(\frac{1}{T}\right) + o_p\left(\frac{1}{|S|}\right)\end{aligned}$$

We then can define $\tilde{\theta}^{MLE}_{N,T(1/2)} := 2\hat{\theta^{MLE}_{N,T}} - \bar{\theta}^{MLE}_{N,T(1/2)}$ where $\bar{\theta}^{MLE}_{N,T(1/2)} := \frac{1}{2}\left(\hat{\theta}^{MLE}_{N,S_1} + \hat{\theta}^{MLE}_{N,S_2}\right)$ where S_1 contains the first $\sim T/2$ elements of the data and S_2 the rest. We then observe that

$$\lim_{N\to\infty} \tilde{\theta}^{MLE}_{N,T(1/2)} = \theta^0 + o_p\left(\frac{1}{T}\right)$$

so that $\sqrt{NT}(\tilde{\theta}^{MLE}_{N,T(1/2)} - \theta^0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$ as $N, T \to \infty$ **WHY IS THIS TRUE?**. This is significant since to first order we have a correctly centered estimator and haven't increased the asymptotic variance of the estimator.

REFERENCES

Manuel Arellano. Dynamic panel data models i: Covariance structures and autoregressions (class notes). Online PDF, CEMFI, October 12 2009. Accessed: 2025-04-20. URL: https://www.cemfi.es/~arellano/time-series-panels-class-note.pdf.

Manuel Arellano. Predetermined variables in panel data models (class notes). Online PDF, CEMFI, 2009. Accessed: 2025-04-20. URL: https://www.cemfi.es/~arellano/predetermined-variables-class-note.pdf.

Manuel Arellano. Static panel data models (class notes). Online PDF, CEMFI, October 9 2009. Accessed: 2025-04-20. URL: https://www.cemfi.es/~arellano/static-panels-class-note.pdf.

Manuel Arellano. Linear panels and random coefficients (cemfi slides). Online PDF, CEMFI, 2017. Accessed: 2025-04-20. URL: https://www.cemfi.es/~arellano/linear-panels-Cemfi-slides-2017.pdf.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009. URL: https://www.sciencedirect.com/science/article/pii/S0304407609000335, doi:10.1016/j.jeconom.2008.12.021.

Charles F. Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993. URL: http://www.jstor.org/stable/2298123.