EM II/QSM II: Development
Problem Set 1 Solutions

Barcelona School of Economics, Spring 2026

Due date: Monday, 19 January, 23:59

Group members: AB, CD, EF, GH.

January 15, 2026

Submission Instructions

Please submit a document with your answers, including Stata or R output and programs, to our TA Janik Deutscher via the Google Classroom by Monday, 19 January, 23:59 at the very latest. You can work in groups of up to four.

1 Random Effects Model

Consider the model with a single regressor x_{it}:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}\,, \tag{1}$$

where α_i represents an unobserved effect fixed over time and u_{it} is a homoskedastic error term which is independent over time t and individuals i. There are N randomly sampled individuals, each observed for $T = 4$ time periods. Assume that $\mathbb{E}(u_{it}|X_i, \alpha_i) = 0$ for all i and that $\mathbb{E}(u_{it}u_{is}|X_i, \alpha_i) = 0$ for any t and $s : t \neq s$ where X_i represents the $T \times 2$ data matrix for individual i.

1. State under which assumptions you would estimate a random effects model in this context. Derive the random effects estimator and show that it is a consistent estimator of $\boldsymbol{\beta} = [\beta_0, \beta_1]'$.

Solution:

Part 1: Assumptions for the Random Effects Model

The random effects model is appropriate under the following assumptions:

RE1. Random individual effects: The unobserved effect is uncorrelated with the regressors:

$$\mathbb{E}(\alpha_i|X_i) = \mathbb{E}(\alpha_i) = 0 \tag{2}$$

where X_i is the $T \times 2$ data matrix for individual i. This is the key distinction from fixed effects, which allows $\mathbb{E}(\alpha_i|X_i) \neq 0$.

RE2. Strictly exogenous regressors: The idiosyncratic error is mean-independent of all regressors (past, present, and future) conditional on the unobserved effect:

$$\mathbb{E}(u_{it}|X_i, \alpha_i) = 0 \quad \text{for all } t = 1, \ldots, T \tag{3}$$

RE3. Homoscedasticity: The variance of both error components is constant:

$$\mathbb{V}\text{ar}(\alpha_i|X_i) = \sigma_\alpha^2 \tag{4}$$

$$\mathbb{V}\text{ar}(u_{it}|X_i, \alpha_i) = \sigma_u^2 \tag{5}$$

$$\mathbb{C}\text{ov}(u_{it}, u_{is}|X_i, \alpha_i) = 0 \quad \text{for } t \neq s \tag{6}$$

RE4. Rank condition: The matrix $\mathbb{E}(X_i'X_i)$ is positive definite with rank 2 (full column rank).

Part 2: Covariance Structure of the Composite Error

Define the composite error $v_{it} = \alpha_i + u_{it}$. Under RE1-RE3, we can derive:

$$\mathbb{C}\text{ov}(v_{it}, v_{is}|X_i) = \mathbb{E}(v_{it}v_{is}|X_i) \tag{7}$$

$$= \mathbb{E}((\alpha_i + u_{it})(\alpha_i + u_{is})|X_i) \tag{8}$$

$$= \mathbb{E}(\alpha_i^2|X_i) + \mathbb{E}(u_{it}u_{is}|X_i) + \mathbb{E}(\alpha_i u_{it}|X_i) + \mathbb{E}(\alpha_i u_{is}|X_i) \tag{9}$$

$$= \sigma_\alpha^2 + \delta_{ts}\sigma_u^2 \tag{10}$$

where δ_{ts} is the Kronecker delta (1 if $t = s$, 0 otherwise). The cross-terms vanish by iterated expectations using RE2.

For individual i, stack the $T = 4$ observations: $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})'$. The covariance matrix is:

$$V = \mathbb{E}(v_i v_i'|X_i) = \sigma_u^2 I_T + \sigma_\alpha^2 J_T \tag{11}$$

where I_T is the $T \times T$ identity matrix and J_T is a $T \times T$ matrix of ones.

Part 3: Derivation of the Random Effects Estimator

The model in vector form for individual i is:

$$y_i = X_i\boldsymbol{\beta} + v_i \tag{12}$$

Since $\mathbb{E}(v_i v_i'|X_i) = V$ is not diagonal, OLS is inefficient (though consistent). The efficient estimator is the Generalized Least Squares (GLS) estimator, which minimizes the weighted sum of squared residuals.

The GLS estimator for the entire sample is:

$$\hat{\boldsymbol{\beta}}^{RE} = \left(\sum_{i=1}^{N} X_i'V^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'V^{-1}y_i\right) \tag{13}$$

Part 4: Proof of Consistency

To show consistency, we demonstrate that $\hat{\boldsymbol{\beta}}^{RE} \xrightarrow{p} \boldsymbol{\beta}$ as $N \to \infty$.

Substituting $y_i = X_i\boldsymbol{\beta} + v_i$ into the estimator:

$$\hat{\boldsymbol{\beta}}^{RE} = \left(\sum_{i=1}^{N} X_i' V^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' V^{-1}(X_i\boldsymbol{\beta} + v_i)\right) \tag{14}$$

$$= \boldsymbol{\beta} + \left(\sum_{i=1}^{N} X_i' V^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' V^{-1} v_i\right) \tag{15}$$

For consistency, we need:

1. $\frac{1}{N}\sum_{i=1}^{N} X_i' V^{-1} X_i \xrightarrow{p} \mathbb{E}(X_i' V^{-1} X_i)$, which is positive definite by RE4.
2. $\frac{1}{N}\sum_{i=1}^{N} X_i' V^{-1} v_i \xrightarrow{p} \mathbb{E}(X_i' V^{-1} v_i)$.

Both follow from the Law of Large Numbers under random sampling. For the second term:

$$\mathbb{E}(X_i' V^{-1} v_i) = \mathbb{E}(\mathbb{E}(X_i' V^{-1} v_i | X_i)) \tag{16}$$

$$= \mathbb{E}(X_i' V^{-1}\, \mathbb{E}(v_i | X_i)) \tag{17}$$

$$= 0 \tag{18}$$

The last equality holds because under RE1 and RE2, $\mathbb{E}(v_i|X_i) = \mathbb{E}(\alpha_i|X_i) + \mathbb{E}(u_i|X_i) = 0$.

Therefore:

$$\hat{\boldsymbol{\beta}}^{RE} - \boldsymbol{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i' V^{-1} X_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} X_i' V^{-1} v_i\right) \xrightarrow{p} \mathbb{E}(X_i' V^{-1} X_i)^{-1} \cdot 0 = 0 \tag{19}$$

Hence, $\hat{\boldsymbol{\beta}}^{RE}$ is a consistent estimator of $\boldsymbol{\beta}$.

2. Derive the (asymptotic) variance-covariance matrix of the random effects estimator.

Solution:

Derivation of the Asymptotic Variance-Covariance Matrix

From Question 1, we have:

$$\hat{\boldsymbol{\beta}}^{RE} - \boldsymbol{\beta} = \left(\sum_{i=1}^{N} X_i' V^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' V^{-1} v_i\right) \tag{20}$$

Multiplying both sides by $\sqrt{N}$:

$$\sqrt{N}(\hat{\boldsymbol{\beta}}^{RE} - \boldsymbol{\beta}) = \left(\frac{1}{N}\sum_{i=1}^{N} X_i' V^{-1} X_i\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i' V^{-1} v_i\right) \tag{21}$$

Step 1: Apply the Law of Large Numbers

Under random sampling and RE4:

$$\frac{1}{N}\sum_{i=1}^{N} X_i' V^{-1} X_i \xrightarrow{p} \mathbb{E}(X_i' V^{-1} X_i) \equiv Q \tag{22}$$

where Q is a positive definite 2×2 matrix.

Step 2: Apply the Central Limit Theorem

Consider the term $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i' V^{-1} v_i$. Define $Z_i = X_i' V^{-1} v_i$, which is a 2×1 random vector. Under random sampling:

- Z_i are i.i.d. across i.
- $\mathbb{E}(Z_i) = \mathbb{E}(X_i' V^{-1} v_i) = 0$ (shown in Question 1).
- $\mathbb{V}\text{ar}(Z_i) = \mathbb{E}(Z_i Z_i') = \mathbb{E}(X_i' V^{-1} v_i v_i' V^{-1} X_i)$.

By the Central Limit Theorem:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} Z_i \xrightarrow{d} \mathcal{N}(0, \Sigma) \tag{23}$$

where $\Sigma = \mathbb{V}\text{ar}(Z_i) = \mathbb{E}(X_i' V^{-1} v_i v_i' V^{-1} X_i)$.

Step 3: Calculate the Variance Σ

Under RE1-RE3, $\mathbb{E}(v_i v_i' | X_i) = V$, so:

$$\begin{aligned}
\Sigma &= \mathbb{E}(X_i' V^{-1} v_i v_i' V^{-1} X_i) && (24)\\
&= \mathbb{E}(\mathbb{E}(X_i' V^{-1} v_i v_i' V^{-1} X_i | X_i)) && (25)\\
&= \mathbb{E}(X_i' V^{-1} \, \mathbb{E}(v_i v_i' | X_i) V^{-1} X_i) && (26)\\
&= \mathbb{E}(X_i' V^{-1} V V^{-1} X_i) && (27)\\
&= \mathbb{E}(X_i' V^{-1} X_i) && (28)\\
&= Q && (29)
\end{aligned}$$

Step 4: Derive the Asymptotic Distribution

Combining Steps 1-3 and applying Slutsky's Theorem:

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\beta}}^{RE} - \boldsymbol{\beta}) &= \left(\frac{1}{N}\sum_{i=1}^{N} X_i' V^{-1} X_i\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i' V^{-1} v_i\right) && (30)\\
&\xrightarrow{d} Q^{-1}\mathcal{N}(0, Q) && (31)\\
&= \mathcal{N}(0, Q^{-1} Q Q^{-1}) && (32)\\
&= \mathcal{N}(0, Q^{-1}) && (33)
\end{aligned}$$

Step 5: State the Final Result

The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}^{RE}$ is:

$$\text{Avar}(\hat{\boldsymbol{\beta}}^{RE}) = \frac{1}{N}\left(\mathbb{E}(X_i' V^{-1} X_i)\right)^{-1} \tag{34}$$

This can be consistently estimated by:

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}^{RE}) = \left(\sum_{i=1}^{N} X_i' \hat{V}^{-1} X_i\right)^{-1} \tag{35}$$

where $\hat{V}$ is a consistent estimator of V (obtained from the feasible GLS procedure, as discussed in Question 3).

Interpretation: The asymptotic variance depends on:

- The inverse of the variance components (through V^{-1}): smaller variance in the composite error leads to more precise estimates.
- The variation in the regressors: more variation in X_i leads to smaller variance.
- The sample size N: the variance decreases at rate $1/N$.

3. Explain how you would implement this estimator using the data in your sample.

Solution:

Feasible GLS Implementation

Since the variance components σ_α^2 and σ_u^2 are unknown in practice, we cannot directly compute $V = \sigma_u^2 I_T + \sigma_\alpha^2 J_T$. Instead, we use a *feasible GLS* (FGLS) approach, which estimates these parameters from the data and then applies GLS with the estimated covariance matrix.

Step 1: Pooled OLS Estimation

First, estimate the model by pooled OLS, ignoring the panel structure:

$$\hat{\boldsymbol{\beta}}^{POLS} = \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i' y_i\right) \tag{36}$$

Compute the residuals:

$$\hat{v}_{it} = y_{it} - x_{it}' \hat{\boldsymbol{\beta}}^{POLS} \tag{37}$$

Although pooled OLS is inefficient (due to serial correlation in v_{it}), it is consistent under RE1-RE2 and provides consistent estimates of the residuals.

Step 2: Estimate Variance Components

Using the POLS residuals, estimate the variance of the composite error:

$$\hat{\sigma}_v^2 = \frac{1}{NT - K} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{v}_{it}^2 \tag{38}$$

where $K = 2$ is the number of regressors. With $T = 4$, this uses $4N - 2$ degrees of freedom.

Next, estimate the covariance between residuals at different time periods for the same individual:

$$\hat{\sigma}_\alpha^2 = \frac{1}{NT(T-1)/2 - K} \sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{s=1}^{t-1} \hat{v}_{it} \hat{v}_{is} \tag{39}$$

This averages all distinct pairs of time periods. For $T = 4$, there are $\binom{4}{2} = 6$ pairs per individual, giving $6N - 2$ degrees of freedom.

Finally, recover the idiosyncratic variance:

$$\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_\alpha^2 \tag{40}$$

This follows from $\mathbb{V}\text{ar}(v_{it}) = \mathbb{V}\text{ar}(\alpha_i + u_{it}) = \sigma_\alpha^2 + \sigma_u^2$ and $\mathbb{C}\text{ov}(v_{it}, v_{is}) = \sigma_\alpha^2$ for $t \neq s$.

Step 3: Construct the Estimated Covariance Matrix

Construct $\hat{V}$ using the estimated variance components:

$$\hat{V} = \hat{\sigma}_u^2 I_4 + \hat{\sigma}_\alpha^2 J_4 = \begin{pmatrix} \hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 \\ \hat{\sigma}_\alpha^2 & \hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 \\ \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 & \hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 \\ \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 & \hat{\sigma}_\alpha^2 & \hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2 \end{pmatrix} \tag{41}$$

Step 4: Compute Feasible GLS Estimator

Apply GLS using $\hat{V}$:

$$\hat{\boldsymbol{\beta}}^{FGLS} = \left(\sum_{i=1}^{N} X_i' \hat{V}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' \hat{V}^{-1} y_i \right) \tag{42}$$

Alternative Implementation via Transformation

Rather than directly inverting $\hat{V}$, it is computationally more efficient to transform the data and apply OLS. Define:

$$\theta = 1 - \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + T\hat{\sigma}_\alpha^2}} = 1 - \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + 4\hat{\sigma}_\alpha^2}} \tag{43}$$

Transform the data by *quasi-demeaning*:

$$y_{it}^* = y_{it} - \theta \bar{y}_i \tag{44}$$

$$x_{it}^* = x_{it} - \theta \bar{x}_i \tag{45}$$

where $\bar{y}_i = \frac{1}{4}\sum_{t=1}^{4} y_{it}$ and $\bar{x}_i = \frac{1}{4}\sum_{t=1}^{4} x_{it}$ are individual-specific time averages.

Then apply OLS to the transformed data:

$$\hat{\beta}^{FGLS} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^{*'} x_{it}^{*}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^{*'} y_{it}^{*}\right) \tag{46}$$

Interpretation of the Transformation:

- If $\theta = 0$ (i.e., $\hat{\sigma}_{\alpha}^2 = 0$): no transformation, equivalent to pooled OLS.
- If $\theta = 1$ (i.e., $\hat{\sigma}_{u}^2 = 0$): full demeaning, equivalent to fixed effects.
- In general, $0 < \theta < 1$: partial demeaning, balancing between and within variation.

Practical Software Implementation:

In Stata:

```
xtset id time
xtreg y x, re
```

In R (using the plm package):

```
library(plm)
model <- plm(y ~ x, data=panel_data, model="random")
```

Both commands automatically perform all the steps above: estimate variance components from POLS residuals, compute the transformation parameter, and apply FGLS.

2 Fixed Effects Variance Estimation

1. Show in detail why, in the context of the fixed effects model, we need to use the formula

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{u}_{it}^2 \tag{47}$$

to obtain a consistent estimate of $\hat{\sigma}_u^2$ (we are ignoring the degrees of freedom adjustment for the K regressors here which is asymptotically irrelevant). In particular, show that we need to divide by $N(T-1)$ rather than NT. (Note: for simplicity, it is fine here to work with $\tilde{u}_{it}$ rather than $\hat{\tilde{u}}_{it}$).

Solution:

Derivation of the Correct Degrees of Freedom for Fixed Effects Variance Estimation

In the fixed effects model, we eliminate the individual-specific effect f_i by applying the within transformation (demeaning). This transformation has important implications for the degrees of freedom available for variance estimation.

Step 1: Setup and Notation

In the fixed effects model, we transform the original error u_{it} into the demeaned error:

$$\tilde{u}_{it} = u_{it} - \bar{u}_i \tag{48}$$

where $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$ is the individual-specific time average of the errors.

Under the fixed effects assumptions (FE1-FE4), we have:

- $\mathbb{E}(u_{it}|X_i, f_i) = 0$ for all t (strict exogeneity)
- $\mathbb{V}\text{ar}(u_{it}|X_i, f_i) = \sigma_u^2$ (homoscedasticity)
- $\mathbb{C}\text{ov}(u_{it}, u_{is}|X_i, f_i) = 0$ for $t \neq s$ (no serial correlation)

Step 2: Calculate the Expected Value of $\tilde{u}_{it}^2$

To understand why we divide by $N(T-1)$ instead of NT, we need to compute $\mathbb{E}(\tilde{u}_{it}^2|X_i, f_i)$.

Expanding the squared demeaned error:

$$\tilde{u}_{it}^2 = (u_{it} - \bar{u}_i)^2 \tag{49}$$

$$= u_{it}^2 - 2u_{it}\bar{u}_i + \bar{u}_i^2 \tag{50}$$

Taking expectations conditional on X_i and f_i:

$$\mathbb{E}(\tilde{u}_{it}^2|X_i, f_i) = \mathbb{E}(u_{it}^2|X_i, f_i) - 2\,\mathbb{E}(u_{it}\bar{u}_i|X_i, f_i) + \mathbb{E}(\bar{u}_i^2|X_i, f_i) \tag{51}$$

Now evaluate each term:

First term:

$$\mathbb{E}(u_{it}^2|X_i, f_i) = \sigma_u^2 \tag{52}$$

Second term:

$$\mathbb{E}(u_{it}\bar{u}_i|X_i, f_i) = \mathbb{E}(u_{it} \cdot \frac{1}{T}\sum_{s=1}^{T} u_{is}|X_i, f_i) \tag{53}$$

$$= \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}(u_{it}u_{is}|X_i, f_i) \tag{54}$$

$$= \frac{1}{T}\,\mathbb{E}(u_{it}^2|X_i, f_i) \quad \text{(only the } s = t \text{ term survives)} \tag{55}$$

$$= \frac{\sigma_u^2}{T} \tag{56}$$

Third term:

$$\mathbb{E}(\bar{u}_i^2 | X_i, f_i) = \mathbb{E}(\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}\right)^2 | X_i, f_i) \tag{57}$$

$$= \frac{1}{T^2}\,\mathbb{E}(\sum_{t=1}^{T}\sum_{s=1}^{T} u_{it}u_{is} | X_i, f_i) \tag{58}$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}(u_{it}^2 | X_i, f_i) \quad \text{(cross terms vanish)} \tag{59}$$

$$= \frac{1}{T^2}\cdot T\sigma_u^2 \tag{60}$$

$$= \frac{\sigma_u^2}{T} \tag{61}$$

Combining these results:

$$\mathbb{E}(\tilde{u}_{it}^2 | X_i, f_i) = \sigma_u^2 - 2\cdot\frac{\sigma_u^2}{T} + \frac{\sigma_u^2}{T} \tag{62}$$

$$= \sigma_u^2 - \frac{\sigma_u^2}{T} \tag{63}$$

$$= \sigma_u^2\left(1 - \frac{1}{T}\right) \tag{64}$$

$$= \sigma_u^2\cdot\frac{T-1}{T} \tag{65}$$

Step 3: Sum Over All Observations

Summing over all individuals and time periods:

$$\mathbb{E}(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 | X, f) = \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}(\tilde{u}_{it}^2 | X_i, f_i) \tag{66}$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_u^2\cdot\frac{T-1}{T} \tag{67}$$

$$= NT\cdot\sigma_u^2\cdot\frac{T-1}{T} \tag{68}$$

$$= N(T-1)\sigma_u^2 \tag{69}$$

Step 4: Derive the Consistent Estimator

To obtain an unbiased (and consistent) estimator of σ_u^2, we need:

$$\mathbb{E}(\hat{\sigma}_u^2) = \sigma_u^2 \tag{70}$$

From Step 3, we have $\mathbb{E}(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2) = N(T-1)\sigma_u^2$. Therefore:

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \tag{71}$$

is the correct estimator.

Intuition:

The demeaning transformation imposes N linear constraints (one for each individual), as:

$$\sum_{t=1}^{T} \tilde{u}_{it} = \sum_{t=1}^{T} (u_{it} - \bar{u}_i) = 0 \quad \text{for each } i \tag{72}$$

Therefore, out of NT total observations, only $N(T-1)$ are "free" after demeaning. Each individual contributes $(T-1)$ degrees of freedom rather than T. This is why the denominator must be $N(T-1)$ rather than NT.

If we incorrectly used NT in the denominator, we would obtain:

$$\mathbb{E}(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{u}_{it}^2) = \frac{N(T-1)\sigma_u^2}{NT} = \frac{T-1}{T}\sigma_u^2 < \sigma_u^2 \tag{73}$$

which would systematically underestimate the true variance σ_u^2.

3 Fixed Effects versus First Difference Estimator

1. Consider the following estimation equation:

$$y_{it} = \alpha + x_{it}\beta + f_i + u_{it} \tag{74}$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$.

where α is a constant, x_{it} is a single time-varying regressor and the idiosyncratic errors are serially uncorrelated and homoscedastic. Show that if $T = 2$, the fixed effects estimator and first difference estimator (which you obtain from transforming the model to $\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}$ and then applying OLS to the transformed model) lead to identical estimates of both the coefficient and its variance.

Solution:

We will show that when $T = 2$, the fixed effects (FE) and first difference (FD) estimators yield identical coefficient estimates and identical variance estimates.

Part A: Equivalence of Coefficient Estimates

Step 1: The Model with $T = 2$

For each individual i, we have two time periods:

$$y_{i1} = \alpha + x_{i1}\beta + f_i + u_{i1} \tag{75}$$

$$y_{i2} = \alpha + x_{i2}\beta + f_i + u_{i2} \tag{76}$$

Step 2: The Fixed Effects Transformation

The FE estimator applies the within transformation, demeaning each variable:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i \quad \text{where} \quad \bar{y}_i = \frac{1}{2}(y_{i1} + y_{i2}) \tag{77}$$

For $T = 2$, the demeaned values are:

$$\tilde{y}_{i1} = y_{i1} - \frac{1}{2}(y_{i1} + y_{i2}) = \frac{1}{2}(y_{i1} - y_{i2}) = -\frac{1}{2}(y_{i2} - y_{i1}) \tag{78}$$

$$\tilde{y}_{i2} = y_{i2} - \frac{1}{2}(y_{i1} + y_{i2}) = \frac{1}{2}(y_{i2} - y_{i1}) \tag{79}$$

Similarly, for x_{it}:

$$\tilde{x}_{i1} = x_{i1} - \frac{1}{2}(x_{i1} + x_{i2}) = -\frac{1}{2}(x_{i2} - x_{i1}) \tag{80}$$

$$\tilde{x}_{i2} = x_{i2} - \frac{1}{2}(x_{i1} + x_{i2}) = \frac{1}{2}(x_{i2} - x_{i1}) \tag{81}$$

And for u_{it}:

$$\tilde{u}_{i1} = u_{i1} - \frac{1}{2}(u_{i1} + u_{i2}) = -\frac{1}{2}(u_{i2} - u_{i1}) \tag{82}$$

$$\tilde{u}_{i2} = u_{i2} - \frac{1}{2}(u_{i1} + u_{i2}) = \frac{1}{2}(u_{i2} - u_{i1}) \tag{83}$$

Note that both f_i and α drop out after demeaning since they are time-invariant.

The FE estimator is then:

$$\hat{\beta}^{FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{2}\tilde{x}_{it}\tilde{y}_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{2}\tilde{x}_{it}^2} \tag{84}$$

Step 3: Simplify the FE Estimator for $T = 2$

The numerator is:

$$\sum_{i=1}^{N}\sum_{t=1}^{2}\tilde{x}_{it}\tilde{y}_{it} = \sum_{i=1}^{N}(\tilde{x}_{i1}\tilde{y}_{i1} + \tilde{x}_{i2}\tilde{y}_{i2}) \tag{85}$$

$$= \sum_{i=1}^{N}\left[-\frac{1}{2}(x_{i2} - x_{i1}) \cdot (-\frac{1}{2}(y_{i2} - y_{i1})) + \frac{1}{2}(x_{i2} - x_{i1}) \cdot \frac{1}{2}(y_{i2} - y_{i1})\right] \tag{86}$$

$$= \sum_{i=1}^{N}\left[\frac{1}{4}(x_{i2} - x_{i1})(y_{i2} - y_{i1}) + \frac{1}{4}(x_{i2} - x_{i1})(y_{i2} - y_{i1})\right] \tag{87}$$

$$= \sum_{i=1}^{N}\frac{1}{2}(x_{i2} - x_{i1})(y_{i2} - y_{i1}) \tag{88}$$

The denominator is:

$$\sum_{i=1}^{N}\sum_{t=1}^{2}\tilde{x}_{it}^2 = \sum_{i=1}^{N}(\tilde{x}_{i1}^2 + \tilde{x}_{i2}^2) \tag{89}$$

$$= \sum_{i=1}^{N}\left[\frac{1}{4}(x_{i2}-x_{i1})^2 + \frac{1}{4}(x_{i2}-x_{i1})^2\right] \tag{90}$$

$$= \sum_{i=1}^{N}\frac{1}{2}(x_{i2}-x_{i1})^2 \tag{91}$$

Therefore:

$$\hat{\beta}^{FE} = \frac{\sum_{i=1}^{N}\frac{1}{2}(x_{i2}-x_{i1})(y_{i2}-y_{i1})}{\sum_{i=1}^{N}\frac{1}{2}(x_{i2}-x_{i1})^2} = \frac{\sum_{i=1}^{N}(x_{i2}-x_{i1})(y_{i2}-y_{i1})}{\sum_{i=1}^{N}(x_{i2}-x_{i1})^2} \tag{92}$$

Step 4: The First Difference Estimator

The FD estimator takes first differences to eliminate f_i and α. For $T = 2$, we have only one difference per individual:

$$\Delta y_{i2} = y_{i2} - y_{i1} = (x_{i2}-x_{i1})\beta + (u_{i2}-u_{i1}) \tag{93}$$

$$= \Delta x_{i2}\beta + \Delta u_{i2} \tag{94}$$

where $\Delta x_{i2} = x_{i2} - x_{i1}$ and $\Delta u_{i2} = u_{i2} - u_{i1}$.

Applying OLS to the differenced model:

$$\hat{\beta}^{FD} = \frac{\sum_{i=1}^{N}\Delta x_{i2}\Delta y_{i2}}{\sum_{i=1}^{N}(\Delta x_{i2})^2} = \frac{\sum_{i=1}^{N}(x_{i2}-x_{i1})(y_{i2}-y_{i1})}{\sum_{i=1}^{N}(x_{i2}-x_{i1})^2} \tag{95}$$

Step 5: Conclusion for Coefficient Estimates

Comparing the expressions from Steps 3 and 4:

$$\hat{\beta}^{FE} = \frac{\sum_{i=1}^{N}(x_{i2}-x_{i1})(y_{i2}-y_{i1})}{\sum_{i=1}^{N}(x_{i2}-x_{i1})^2} = \hat{\beta}^{FD} \tag{96}$$

Therefore, when $T = 2$, the FE and FD estimators produce identical coefficient estimates.

Part B: Equivalence of Variance Estimates

Step 6: Variance of the FE Estimator

Under homoscedasticity ($\mathbb{V}\text{ar}(u_{it}|X_i, f_i) = \sigma_u^2$) and no serial correlation, the variance of the FE estimator is:

$$\mathbb{V}\text{ar}(\hat{\beta}^{FE}|X, f) = \frac{\sigma_u^2}{\sum_{i=1}^{N}\sum_{t=1}^{2}\tilde{x}_{it}^2} \tag{97}$$

From Step 3, we know that:

$$\sum_{i=1}^{N}\sum_{t=1}^{2}\tilde{x}_{it}^2 = \sum_{i=1}^{N}\frac{1}{2}(x_{i2}-x_{i1})^2 = \frac{1}{2}\sum_{i=1}^{N}(\Delta x_{i2})^2 \tag{98}$$

Therefore:

$$\mathbb{V}\text{ar}(\hat{\beta}^{FE}|X,f) = \frac{\sigma_u^2}{\frac{1}{2}\sum_{i=1}^{N}(\Delta x_{i2})^2} = \frac{2\sigma_u^2}{\sum_{i=1}^{N}(\Delta x_{i2})^2} \tag{99}$$

Step 7: Variance of the FD Estimator

For the FD estimator, the error term is $\Delta u_{i2} = u_{i2} - u_{i1}$. Under the assumptions of homoscedasticity and no serial correlation:

$$\begin{aligned}
\mathbb{V}\text{ar}(\Delta u_{i2}|X_i,f_i) &= \mathbb{V}\text{ar}(u_{i2}-u_{i1}|X_i,f_i) && (100)\\
&= \mathbb{V}\text{ar}(u_{i2}|X_i,f_i) + \mathbb{V}\text{ar}(u_{i1}|X_i,f_i) - 2\,\mathbb{C}\text{ov}(u_{i2},u_{i1}|X_i,f_i) && (101)\\
&= \sigma_u^2 + \sigma_u^2 - 0 && (102)\\
&= 2\sigma_u^2 && (103)
\end{aligned}$$

The variance of the FD estimator is:

$$\mathbb{V}\text{ar}(\hat{\beta}^{FD}|X,f) = \frac{\mathbb{V}\text{ar}(\Delta u_{i2})}{\sum_{i=1}^{N}(\Delta x_{i2})^2} = \frac{2\sigma_u^2}{\sum_{i=1}^{N}(\Delta x_{i2})^2} \tag{104}$$

Step 8: Conclusion for Variance Estimates

Comparing the variances from Steps 6 and 7:

$$\mathbb{V}\text{ar}(\hat{\beta}^{FE}|X,f) = \frac{2\sigma_u^2}{\sum_{i=1}^{N}(\Delta x_{i2})^2} = \mathbb{V}\text{ar}(\hat{\beta}^{FD}|X,f) \tag{105}$$

Therefore, when $T = 2$, the FE and FD estimators also have identical variances.

Summary:

When $T = 2$, the within transformation and first differencing are algebraically equivalent transformations. The FE estimator demeans the data, which with two periods amounts to taking half the difference. The FD estimator takes the full difference. Since both numerators and denominators differ by the same factor of 2, the coefficient estimates are identical. Moreover, because the variance of the differenced error is exactly twice the variance of the original error, and the denominator of the FE variance is exactly half that of the FD variance, the two variance formulas are also identical.

4 Empirical Analysis: Children and Life Satisfaction

In this question, we will use panel data methods to understand how having children affects life satisfaction. Download the SOEP practise data set `soep_lebensz_en.dta`, which is available on the course website, and inspect its variables (using Stata or R).

1. Construct a binary variable `has_kids` that indicates if a person at time t has any children at all. Then, regress the (standardized) variable measuring current life satisfaction, `satisf_std`, on your constructed indicator. Include the individual's gender, education, categorical health and indicator variables for each year in the regression, and cluster your standard errors at the level of the individual.

   First, estimate the effect of the children indicator on life satisfaction in a pooled OLS regression. Then, estimate the effect with a fixed effects regression. What does the difference of the estimated coefficients tell you about the unobserved effect f_i and, in particular, its covariance with `has_kids`?

   Solution: Here comes the solution.

2. Why has the coefficient on gender disappeared in the fixed effects regression? Run the same fixed effects regression as in the previous question but this time interact the gender indicator with the children indicator. Are women and men affected differentially? How do you interpret the magnitudes of the estimated coefficients?

   Solution: Here comes the solution.

3. Test the effect of having children on life satisfaction in a random effects model. Do the coefficients of the children indicator differ between the fixed and random effects model? What can you infer from this? Can you trust the assumptions of the RE model in this context? Why? Why not?

   Solution: Here comes the solution.

4. Perform a formal Hausman test to compare the fixed effects and the random effects model. Do you reject the null hypothesis? What does this result tell you? (Hint: check out the command `hausman` in Stata or `phtest` in R).

   Solution: Here comes the solution.