EM II/QSM II: Development
Problem Set 2 Solutions

Barcelona School of Economics, Spring 2026

Due date: Thursday, 29 January, 18:00

Group members: AB, CD, EF, GH.

January 22, 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 Thursday, 29 January, 18:00 at the very latest. You can work in groups of up to four.

1 Pooled OLS Estimation

You have a sample of N individuals for T years. Suppose you estimate by Pooled OLS the annual income equation:

$$Y_{it} = \alpha_0 + \alpha_1 ed_i + \alpha_2 age_{it} + \alpha_3 (ed_i \times age_{it}) + \gamma Y_{it-1} + u_{it}\,, \tag{1}$$

where ed_i represents the years of education of the ith individual, age_{it} represents the age of the individual i in period t, and u_{it} represents all unobservables.

1. Suppose you estimate γ as 0.82 with a standard error of 0.05. State a set of sufficient assumptions for the consistency of the Pooled OLS estimator in this context.

Solution: For consistency of Pooled OLS in this dynamic panel model with lagged dependent variable Y_{it-1}, we require:

Assumption 1: Random Effects. Decompose $u_{it} = f_i + \varepsilon_{it}$ where f_i is a time-invariant individual effect and ε_{it} is the idiosyncratic error. Then:

$$\mathbb{E}(f_i | ed_i, age_{i1}, \ldots, age_{iT}, Y_{i0}) = 0\,. \tag{2}$$

This ensures Y_{it-1} is uncorrelated with f_i.

Assumption 2: Sequential exogeneity. The idiosyncratic error is mean independent of past outcomes and current/past regressors:

$$\mathbb{E}(\varepsilon_{it} | ed_i, age_{i1}, \ldots, age_{it}, Y_{i0}, \ldots, Y_{it-1}) = 0\,. \tag{3}$$

Assumption 3: No serial correlation. The idiosyncratic errors are serially uncorrelated:

$$\mathbb{E}(\varepsilon_{it}\varepsilon_{is}) = 0 \quad \text{for all } t \neq s\,. \tag{4}$$

Otherwise Y_{it-1} correlates with ε_{it} through ε_{it-1}.

Assumption 4: Standard regularity conditions. The model is correctly specified and $\mathbb{E}(X'X)$ is positive definite. The initial value Y_{i0} is uncorrelated with $\varepsilon_{i1}, \ldots, \varepsilon_{iT}$.

2. Describe an alternative estimation technique and general procedure that you could use to evaluate the validity of some of your assumptions. Justify your choice and explain carefully the conditions under which your alternative estimator is consistent.

Solution: The **Arellano-Bond GMM estimator** allows us to test the Random Effects assumption by eliminating f_i through first-differencing while addressing endogeneity of Y_{it-1} using lagged levels as instruments.

Procedure:

Step 1: First-difference to eliminate f_i:

$$\Delta Y_{it} = \alpha_2 \Delta age_{it} + \alpha_3 ed_i \Delta age_{it} + \gamma \Delta Y_{it-1} + \Delta u_{it}\,. \tag{5}$$

Step 2: Since ΔY_{it-1} is correlated with Δu_{it} through u_{it-1}, use $Y_{it-2}, Y_{it-3}, \ldots, Y_{i1}$ as instruments for the equation in period t, exploiting:

$$\mathbb{E}(Y_{it-s}\Delta u_{it}) = 0 \quad \text{for } s \geq 2\,. \tag{6}$$

Step 3: Estimate via GMM using all available moment conditions with optimal weighting matrix.

Consistency conditions:

(i) No serial correlation: $\mathbb{E}(u_{it}u_{is}) = 0$ for $t \neq s$. Otherwise Y_{it-2} correlates with u_{it-1} in Δu_{it}, invalidating instruments.

(ii) Strict exogeneity of age_{it}: $\mathbb{E}(u_{it}|age_{i1}, \ldots, age_{iT}, f_i) = 0$.

(iii) Arbitrary correlation allowed: $\mathbb{E}(f_i|ed_i, age_{i1}, \ldots, age_{iT})$ can be non-zero.

(iv) Valid initial conditions: $\mathbb{E}(Y_{i0}u_{it}) = 0$ for $t \geq 2$.

Diagnostic tests:

1. AR(1) test: Should reject (mechanical in differences).
2. AR(2) test: Should not reject (validates no serial correlation in levels).
3. Sargan/Hansen test: Tests overidentifying restrictions.

4. Coefficient comparison: Arellano-Bond $\hat{\gamma}$ typically lies between Pooled OLS (upward biased) and Fixed Effects (Nickell bias).

A significant difference between Pooled OLS and Arellano-Bond estimates indicates violation of the Random Effects assumption.

2 First Differences and Efficient Estimation

Consider the following model:

$$y_{it} = \alpha + x_{it}\beta + z_{it}\gamma + f_i + u_{it} \tag{7}$$

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

where x_{it} and z_{it} are scalars, f_i is a permanent unobserved effect and the error term u_{it} is homoscedastic and serially uncorrelated. Furthermore, assume that x_{it} is strictly exogenous:

$$\mathbb{E}(u_{it}|x_{i1}, \ldots, x_{iT}, f_i) = 0 \quad (1) \tag{8}$$

1. Suppose the only thing we can safely assume is that the orthogonality condition (1) holds and you take first differences to estimate the model. For different assumptions regarding the exogeneity of z_{it} and the relationship between z_{it} and f_i, state the properties of your estimator (consistency and efficiency) when OLS is used to estimate β and γ in the first differences model.

Solution:

First differencing eliminates f_i and α:

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta z_{it}\gamma + \Delta u_{it}\,. \tag{9}$$

Since $\mathbb{E}(u_{it}|x_{i1}, \ldots, x_{iT}, f_i) = 0$, we have $\mathbb{E}(\Delta x_{it}\Delta u_{it}) = 0$. The properties of OLS for β and γ depend on assumptions about z_{it}:

Case 1: z_{it} **strictly exogenous,** $\mathbb{E}(f_i|z_{i1}, \ldots, z_{iT}) = 0$**.** Under $\mathbb{E}(u_{it}|z_{i1}, \ldots, z_{iT}, f_i) = 0$, we have $\mathbb{E}(\Delta z_{it}\Delta u_{it}) = 0$. Consistent: Yes. Efficient: Yes (OLS is BLUE for homoscedastic errors across individuals).

Case 2: z_{it} **strictly exogenous,** $\mathbb{E}(f_i|z_{i1}, \ldots, z_{iT}) \neq 0$**.** Under $\mathbb{E}(u_{it}|z_{i1}, \ldots, z_{iT}, f_i) = 0$, we have $\mathbb{E}(\Delta z_{it}\Delta u_{it}) = 0$. Consistent: Yes. Efficient: Yes. Note: Cases 1 and 2 yield identical properties because first differencing eliminates f_i, rendering any correlation between z_{it} and f_i irrelevant.

Case 3: z_{it} **predetermined (sequential exogeneity).**

Assume $\mathbb{E}(u_{it}|z_{i1}, \ldots, z_{it}, f_i) = 0$ but $\mathbb{E}(u_{it}|z_{it+1}, \ldots, z_{iT}) \neq 0$. Then:

$$\mathbb{E}(\Delta z_{it}\Delta u_{it}) = \mathbb{E}(z_{it}u_{it}) - \mathbb{E}(z_{it}u_{it-1}) - \mathbb{E}(z_{it-1}u_{it}) + \mathbb{E}(z_{it-1}u_{it-1}) = 0\,. \tag{10}$$

Consistent: Yes. Efficient: No (GMM using instruments $z_{i1}, \ldots, z_{it-1}$ would be more efficient).

Case 4: z_{it} **endogenous.** If $\mathbb{E}(u_{it}|z_{it}) \neq 0$, then $\mathbb{E}(\Delta z_{it}\Delta u_{it}) \neq 0$. Consistent: No. Efficiency: N/A.

Assumption on z_{it}	Consistent	Efficient
Strictly exogenous, $\mathbb{E}(f_i|z_i) = 0$	Yes	Yes
Strictly exogenous, $\mathbb{E}(f_i|z_i) \neq 0$	Yes	Yes
Predetermined (sequential exogeneity)	Yes	No
Endogenous	No	N/A

2. Now suppose $T = 5$ and the additional assumption holds:

$$\mathbb{E}(u_{it}|z_{i1}, \ldots, z_{it-1}, f_i) = 0\,. \tag{11}$$

(a) How would you now estimate β and γ efficiently?

Solution:

Under sequential exogeneity $\mathbb{E}(u_{it}|z_{i1}, \ldots, z_{it-1}, f_i) = 0$, estimate efficiently using GMM with all available moment conditions.

Step 1: First-difference to eliminate f_i:

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta z_{it}\gamma + \Delta u_{it} \quad \text{for } t = 2, \ldots, 5\,. \tag{12}$$

Step 2: Construct instrument matrix. For individual i, stack differenced equations and use instruments: all x_{it} values (strict exogeneity) and past z values (sequential exogeneity):

$$Z_i = \begin{pmatrix} x_{i1} & x_{i2} & 0 & 0 & 0 & z_{i1} & 0 & 0 & 0 \\ x_{i1} & x_{i2} & x_{i3} & 0 & 0 & z_{i1} & z_{i2} & 0 & 0 \\ x_{i1} & x_{i2} & x_{i3} & x_{i4} & 0 & z_{i1} & z_{i2} & z_{i3} & 0 \\ x_{i1} & x_{i2} & x_{i3} & x_{i4} & x_{i5} & z_{i1} & z_{i2} & z_{i3} & z_{i4} \end{pmatrix}. \tag{13}$$

This gives $q = 9$ moment conditions $\mathbb{E}(Z_i'\Delta u_i) = 0$ to estimate $k = 2$ parameters (overidentified).

Step 3: The efficient GMM estimator uses optimal weighting matrix $W = \hat{\Omega}^{-1}$:

$$\hat{\theta} = \left(\sum_{i=1}^{N} \Delta X_i' Z_i W Z_i' \Delta X_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta X_i' Z_i W Z_i' \Delta y_i\right), \tag{14}$$

where $\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} Z_i'\hat{\Delta u}_i\hat{\Delta u}_i' Z_i$ using first-stage residuals.

Since u_{it} is serially uncorrelated, Δu_{it} exhibits mechanical serial correlation:

$$\mathbb{E}(\Delta u_{it}\Delta u_{it-1}) = -\sigma_u^2\,. \tag{15}$$

The covariance matrix is:

$$\mathbb{E}(\Delta u_i \Delta u_i') = \sigma_u^2 \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}. \tag{16}$$

Implement via two-step GMM: (i) obtain consistent first-stage estimate, (ii) construct $\hat{\Omega}$, (iii) re-estimate with $W = \hat{\Omega}^{-1}$.

(b) Derive the variance of your estimator.

Solution:

For the GMM estimator with moment conditions $\mathbb{E}(Z_i' \Delta u_i) = 0$, the general asymptotic variance is:

$$\mathbb{V}\text{ar}(()\hat{\theta}) = \frac{1}{N}(G'WG)^{-1}(G'W\Omega WG)(G'WG)^{-1}, \tag{17}$$

where $G = \mathbb{E}(Z_i' \Delta X_i)$ is the $q \times k$ matrix of instruments times regressors, W is the weighting matrix, and $\Omega = \mathbb{E}(Z_i' \Delta u_i \Delta u_i' Z_i)$.

With optimal weighting $W = \Omega^{-1}$, this simplifies to:

$$\boxed{\mathbb{V}\text{ar}(()\hat{\theta}) = \frac{1}{N}(G'\Omega^{-1}G)^{-1} = \frac{1}{N}\left(\mathbb{E}(Z_i'\Delta X_i)' \left[\mathbb{E}(Z_i'\Delta u_i \Delta u_i' Z_i)\right]^{-1} \mathbb{E}(Z_i'\Delta X_i)\right)^{-1}} \tag{18}$$

In practice, estimate by replacing population moments with sample analogs:

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{N}\left(\hat{G}'\hat{\Omega}^{-1}\hat{G}\right)^{-1}, \tag{19}$$

where $\hat{G} = \frac{1}{N}\sum_{i=1}^{N} Z_i' \Delta X_i$ and $\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} Z_i' \hat{\Delta u}_i \hat{\Delta u}_i' Z_i$ using residuals $\hat{\Delta u}_i = \Delta y_i - \Delta X_i \hat{\theta}$.

3. Suppose $z_{it} = y_{it-1}$ and you know that $\gamma = 1$. How would you estimate the model? Provide the estimator and its variance matrix.

Solution:

With $z_{it} = y_{it-1}$ and $\gamma = 1$, the model becomes $y_{it} = \alpha + x_{it}\beta + y_{it-1} + f_i + u_{it}$. First differencing eliminates f_i and α:

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta y_{it-1} + \Delta u_{it}. \tag{20}$$

Since Δy_{it-1} is correlated with Δu_{it} through u_{it-1}, OLS is inconsistent. Use GMM with instruments.

Valid instruments (assuming u_{it} is serially uncorrelated):

- Levels of y lagged 2+ periods: $y_{it-2}, y_{it-3}, \ldots, y_{i1}$ (satisfy $\mathbb{E}(y_{it-s}\Delta u_{it}) = 0$ for $s \geq 2$)

- All values of x_{it}: $x_{i1}, \ldots, x_{iT}$ (strict exogeneity)

Instrument matrix:

$$Z_i = \begin{pmatrix} x_{i1} & x_{i2} & x_{i3} & \cdots & x_{iT} & y_{i1} & 0 & 0 & \cdots \\ x_{i1} & x_{i2} & x_{i3} & \cdots & x_{iT} & y_{i1} & y_{i2} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}. \tag{21}$$

Efficient GMM estimator:

$$\hat{\beta} = \left(\sum_{i=1}^{N} \Delta X_i' Z_i \hat{\Omega}^{-1} Z_i' \Delta X_i \right)^{-1} \left(\sum_{i=1}^{N} \Delta X_i' Z_i \hat{\Omega}^{-1} Z_i' \Delta y_i \right), \tag{22}$$

where ΔX_i contains $[\Delta x_{it}, \Delta y_{it-1}]$ and $\hat{\Omega} = \frac{1}{N} \sum_{i=1}^{N} Z_i' \hat{\Delta u}_i \hat{\Delta u}_i' Z_i$ using first-stage residuals.

Variance matrix:

$$\boxed{\mathbb{V}\text{ar}(()\hat{\beta}) = \frac{1}{N} \left(\mathbb{E}(\Delta X_i' Z_i)' \left[\mathbb{E}(Z_i' \Delta u_i \Delta u_i' Z_i) \right]^{-1} \mathbb{E}(\Delta X_i' Z_i) \right)^{-1}} \tag{23}$$

Estimated by:

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{N} \left(\frac{1}{N} \sum_{i=1}^{N} \Delta X_i' Z_i \hat{\Omega}^{-1} Z_i' \Delta X_i \right)^{-1}. \tag{24}$$

Since u_{it} is serially uncorrelated, $\mathbb{E}(\Delta u_i \Delta u_i') = \sigma_u^2 \Lambda$ where Λ is the tridiagonal matrix with 2 on diagonal and -1 on off-diagonals.

3 Empirical Analysis: Life Satisfaction Dynamics

Download the SOEP practise data set `soep_lebensz_en.dta`, available on the course website. As in Problem Set 1, construct the binary variable `has_kids` that indicates if a person at time t has any children. For each individual, keep only the first two time periods. In the following regressions, include as control variables only education and an individual's current standardized health status, and use non-clustered, non-robust standard errors for simplicity.

1. Estimate the effect of the child indicator on standardized life satisfaction in a first-difference model. Next, estimate the effect with a fixed effects regression. Do you expect the estimated coefficients to differ? How would you interpret the estimated coefficients that you obtain? (Hint: Make sure that you explicitly exclude all variables that are differenced-out in the FD model.)

Solution:

Using the first two time periods for each individual (N=2,654 individuals, 4,980 observations):

	First Difference	Fixed Effects
health_std	0.224***	0.229***
	(0.023)	(0.023)
has_kids	0.022	0.055
	(0.102)	(0.102)
N (individuals)	2,654	2,654
N (observations)	4,980	4,980

*** p<0.001, ** p<0.01, * p<0.05
Standard errors in parentheses.

Expected difference: No. With $T = 2$, FD and FE transformations are algebraically equivalent in a balanced panel. Results confirm this—coefficients are nearly identical, with minor differences due to imperfect balance.

Interpretation: A one standard deviation increase in health increases life satisfaction by 0.22–0.23 standard deviations (highly significant, $p < 0.001$). Having children shows a small positive association (0.02–0.05 SD) but is not statistically significant ($p > 0.5$). Education is time-invariant and drops out in both transformations.

2. Start again with the full sample. This time keep all time periods. Re-estimate the FE and FD specifications of the previous question. Do the estimated coefficients differ? Why? Now assume that the assumptions for consistency of the FE and FD estimators hold in your model. In theory, when is the FD estimator efficient, when the FE estimator? In your context, which estimator would you expect to be more efficient? Why? Which estimate, $\hat{\beta}^{FD}$ or $\hat{\beta}^{FE}$, has the higher standard error? How could you make your standard errors more robust to deviations from the assumed structure of the idiosyncratic error terms?

Solution:

Using the full sample (N=3,289, 10,659 observations):

	Two Periods		Full Sample	
	FD	FE	FD	FE
health_std	0.224***	0.229***	0.207***	0.248***
	(0.023)	(0.023)	(0.013)	(0.013)
has_kids	0.022	0.055	0.058	0.129**
	(0.102)	(0.102)	(0.059)	(0.047)
N (individuals)	2,654	2,654	3,289	3,289
N (observations)	4,980	4,980	10,659	10,659

*** p<0.001, ** p<0.01, * p<0.05
Standard errors in parentheses.

Do coefficients differ? Yes. The health effect remains stable (0.21–0.25), while the has_kids effect increases substantially in FE (0.055 $\rightarrow$ 0.129, now significant at $p = 0.006$). Standard errors are much smaller in the full sample due to greater within-individual variation and larger sample size.

Efficiency theory: FD is efficient when idiosyncratic errors follow a random walk (highly serially correlated); FE is efficient when u_{it} is serially uncorrelated. Life satisfaction likely responds to both persistent factors and transitory shocks, suggesting FE should be more efficient.

Evidence: For has_kids, $SE^{FD} = 0.059$ vs. $SE^{FE} = 0.047$. FE has lower standard error, confirming it is more efficient and suggesting errors are not highly serially correlated.

Robust standard errors: Use cluster-robust standard errors at the individual level to account for arbitrary serial correlation and heteroscedasticity, or Arellano/Driscoll-Kraay standard errors for panel data.

3. Next, we build a dynamic model of life satisfaction. Intuitively, would you expect current life satisfaction and past life satisfaction to be positively or negatively related? Why? Now estimate a fixed effects model that includes, besides the first lag of life satisfaction, an individual's education and current standardized health status as well as an indicator for having children as additional control variables. Which sign has the coefficient of lagged life satisfaction? Is this what you expected? Do you think the coefficient is unbiased? Why? Why not?

Solution:

Expected relationship: We expect a positive relationship due to state dependence, persistent personality traits, and stable life circumstances creating autocorrelation in satisfaction.

Dynamic FE results:

Variable	Coefficient (SE)
lag(satisf_std)	-0.146^{***}
	(0.014)
health_std	0.219^{***}
	(0.013)
education	0.011
	(0.013)
has_kids	0.180^{**}
	(0.049)
N (observations)	7,202

*** p<0.001, ** p<0.01, * p<0.05

Observed sign: Negative (-0.146, $p < 0.001$), contradicting our expectation.

Is the coefficient unbiased? No. This is **Nickell bias**. The within transformation demeans all variables: $\tilde{y}_{it-1} = y_{it-1} - \bar{y}_i$ contains y_{it-1} in both numerator and denominator, while $\tilde{u}_{it} = u_{it} - \bar{u}_i$

contains u_{it-1}. This creates correlation $\mathbb{E}(\tilde{y}_{it-1} \cdot \tilde{u}_{it}) \neq 0$ even when original errors are uncorrelated. The bias is $O(1/T)$ and downward. With average $T \approx 3$–4, this bias is substantial. The observed negative coefficient likely reflects bias rather than true dynamics, requiring an alternative estimator like Arellano-Bond.

4. Now, estimate your dynamic panel data model using the Arellano-Bond Estimator (Hint: use the Stata command `xtabond` or the R package `pdynmc`). Include one lag of the dependent variable and use at most 2 lags as instruments. Under which assumptions is your estimate of the coefficient on the lagged dependent variable unbiased? Test for serial correlation in the error terms u_{it}. What do you conclude? If unbiased, would you expect the Arellano-Bond estimate to be more positive or more negative relative to the FE estimate from the previous question? Why? Which sign has the coefficient of lagged life satisfaction now? Interpret the magnitudes of your estimated coefficients.

Solution:

Arellano-Bond difference GMM using lags 2–3 as instruments (N=4,304):

Variable	Coefficient (SE)
lag(satisf_std)	0.067*
	(0.035)
health_std	0.205***
	(0.014)
education	−0.049
	(0.060)
has_kids	0.114
	(0.075)
N (observations)	4,304
Diagnostic Tests:	
AR(1) test (p-value)	< 0.001
AR(2) test (p-value)	0.363
Sargan test (p-value)	0.223

*** p<0.001, ** p<0.01, * p<0.05

Consistency assumptions: (i) No serial correlation in level errors: $\mathbb{E}(u_{it} \cdot u_{is}) = 0$ for $t \neq s$; (ii) Valid instruments: $\mathbb{E}(y_{it-s} \cdot \Delta u_{it}) = 0$ for $s \geq 2$; (iii) Strict exogeneity (or predetermined) of regressors; (iv) f_i can be correlated with regressors.

Serial correlation tests: AR(1) rejects ($p < 0.001$)—expected and acceptable, arising mechanically from first-differencing when level errors are uncorrelated. AR(2) fails to reject ($p = 0.363$)—this is the key test, confirming no second-order serial correlation in differenced errors, which validates that original level errors are uncorrelated and instruments are valid. Sargan test ($p = 0.223$) supports instrument validity.

Expected sign relative to FE: More positive. FE suffers downward Nickell bias; Arellano-Bond corrects this using valid instruments.

Estimator	Coefficient on lag(satisf_std)
FE Dynamic	-0.146*** (biased downward)
Arellano-Bond	$+0.067$* (consistent)

Arellano-Bond is more positive ($+0.067$ vs. -0.146), correcting 0.21 of downward bias.

Coefficient interpretations: Lagged satisfaction (0.067): weak positive persistence.

Health (0.205***): dominant determinant. Has children (0.114): positive but not significant.

Education (-0.049): not significant.