EI035 Econometrics II
The Graduate Institute, Spring 2025, Marko Mlikota

PS2 Solutions

Jingle Fu

Solution (a).

- The dataset (GMdata.dta) contains firm-level panel data for the years 1973, 1978, 1983, and 1988.
- Key variables include `ldsal`, `lemp`, and `ldnpt` among others.
- After loading the data, initial inspection confirms there are 2,971 firm-year observations.

```
use GMdata.dta, clear

// Compute summary statistics by year for ldsal, lemp, and ldnpt
tabstat ldsal lemp ldnpt, statistics(mean median sd min max p5 p95) by(
    yr)

// Box plot for ldsal (log of deflated sales)
graph box ldsal, over(yr) title("Boxplot of ldsal by Year")
graph export "boxplot_ldsal.png", replace

// Box plot for lemp (log of employment)
graph box lemp, over(yr) title("Boxplot of lemp by Year")
graph export "boxplot_lemp.png", replace

// Box plot for ldnpt (log of deflated capital)
graph box ldnpt, over(yr) title("Boxplot of ldnpt by Year")
graph export "boxplot_ldnpt.png", replace
```

Solution (b).

- **Summary Statistics by Year:** For each year, detailed summaries (mean, median, standard deviation, minimum, maximum, 5th percentile, and 95th percentile) are computed for `ldsal`, `lemp`, and `ldnpt`.
- **Findings:**
  - *1973:* Average log-sales around 6.00, with a median of 5.91; labor and capital show relatively high values.

  - *1978:* A slight dip is observed in log-sales, and both labor and capital statistics indicate a reduction.
  - *1983:* Similar levels to 1978 with very low values at the 5th percentile for capital.
  - *1988:* A modest rebound in log-sales is observed along with a continuation of the decline in typical labor and capital values, although high outliers emerge.

- **Box Plots:** Side-by-side box plots for each variable across the years reveal:
  - For `ldsal`: Median sales dip in the late 1970s and early 1980s then rise by 1988, with an increasing number of high-value outliers.
  - For `lemp`: A downward trend in median employment, indicating that the typical firm employs fewer workers in later years.
  - For `ldnpt`: A similar downward drift in capital, with consistent high-capital outliers.

- **Conclusion:** There is no clear monotonic trend in output; however, the distributions of labor and capital suggest a shift towards smaller firms in later years, while a few firms experienced very high sales by 1988.

```
egen tag = tag(index)
count if tag == 1
local total_firms = r(N)
display "Total number of firms in the unbalanced panel: " `total_firms'

// Step 2: For each firm, count the number of observations (years)
    available.
bysort index: egen n_obs = count(yr)

// Optional: List firms with fewer than 4 observations
list index n_obs if n_obs < 4, sepby(index)

// Step 3: Create a balanced panel by keeping only firms with
    observations in all four years.
keep if n_obs == 4

// Step 4: Count the number of unique firms in the balanced panel.
egen tag_bal = tag(index)
count if tag_bal == 1
local balanced_firms = r(N)
display "Number of firms in the balanced panel: " `balanced_firms'

// Step 5: Calculate and display the number of firms dropped.
local lost_firms = `total_firms' - `balanced_firms'
display "Number of firms dropped when creating a balanced panel: " `
    lost_firms'
```

Solution (c).
Consider the panel data model

$$\text{ldsal}_{it} = \alpha_i + \beta_1 \,\text{lemp}_{it} + \beta_2 \,\text{ldnpt}_{it} + u_{it}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T,$$

where α_i is an unobserved firm-specific effect and u_{it} is an idiosyncratic error term. In the random effects (RE) framework, α_i is treated as a random variable that is uncorrelated with the regressors lemp_{it} and ldnpt_{it}.

1. Model Reformulation

Define the composite error term as

$$\varepsilon_{it} = \alpha_i + u_{it} - \beta_0.$$

Thus, the model can be rewritten as

$$\text{ldsal}_{it} = \beta_0 + \beta_1 \,\text{lemp}_{it} + \beta_2 \,\text{ldnpt}_{it} + \varepsilon_{it},$$

with β_0 being an overall intercept (if included).

2. Error Structure

Under RE, the variance of the composite error is:

$$\text{Var}(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_u^2,$$

and for two different time periods $t \neq s$ for the same firm, the covariance is:

$$\text{Cov}(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2.$$

This intra-firm correlation suggests the use of Generalized Least Squares (GLS).

3. Quasi-Demeaning Transformation

To account for the correlation in the error structure, a quasi-demeaning transformation is applied. For each firm i, define the time averages:

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} \text{ldsal}_{it}, \quad \bar{x}_i = \frac{1}{T} \sum_{t=1}^{T} x_{it},$$

where x_{it} represents each regressor (i.e., lemp_{it} and ldnpt_{it}).

The transformed variables are:

$$y_{it}^* = \text{ldsal}_{it} - \theta \,\bar{y}_i, \quad x_{it}^* = x_{it} - \theta \,\bar{x}_i,$$

with the transformation parameter θ defined as:

$$\theta = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}}.$$

After this transformation, the model becomes:

$$y_{it}^* = \beta_0(1-\theta) + \beta_1 \,\text{lemp}_{it}^* + \beta_2 \,\text{ldnpt}_{it}^* + \varepsilon_{it}^*.$$

Applying ordinary least squares (OLS) to the transformed model yields the RE estimator:

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}^* x_{it}^{*\prime} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}^* y_{it}^* \right).$$

4. Assumptions for Consistency

For the RE estimator to be consistent, the following assumptions must hold:

1. **Exogeneity:**

$$E\left[u_{it} \mid \text{lemp}_{i1}, \text{ldnpt}_{i1}, \ldots, \text{lemp}_{iT}, \text{ldnpt}_{iT}, \alpha_i\right] = 0.$$

   Equivalently, the regressors are uncorrelated with both the idiosyncratic error u_{it} and the individual effect α_i.

2. **Independence of Individual Effects and Regressors:**

$$E\left[\alpha_i \mid \text{lemp}_{it}, \text{ldnpt}_{it}\right] = 0 \quad \text{for all } t.$$

3. **Homoskedasticity and No Serial Correlation:** The idiosyncratic errors u_{it} are assumed to be homoskedastic and serially uncorrelated (conditional on α_i).

4. **Correct Specification:** The model is correctly specified with no omitted variables that are correlated with the regressors.

5. Discussion

In many empirical applications such as in production function estimation (e.g., Griliches and Mairesse, 1995), the assumption that α_i is uncorrelated with the inputs might be unrealistic. If there exists correlation between α_i and the regressors, the RE estimator will be inconsistent, and alternative methods like the Fixed Effects estimator should be considered.

- The RE estimator is obtained by running a pooled OLS regression where the firm-specific effect α_i is absorbed into the error term.

- **Results (balanced panel):**

Table 1: Question (c): Random Effects Estimator

	(1)		(2)	
	RE		RE Robust	
	b	se	b	se
log of employment	0.509	0.037	0.509	0.041
log of deflated capital	0.484	0.029	0.484	0.035
Observations	856		856	

- Intercept: approximately 2.724.
- β_1 (Labor elasticity): about 0.509.
- β_2 (Capital elasticity): about 0.484.

- **Interpretation:** A 1% increase in labor is associated with a 0.5087% increase in sales, and a 1% increase in capital with a 0.4839% increase in sales. The sum (approximately 0.989) is close to constant returns.
- **Assumptions:** Consistency requires that the unobserved firm effect α_i is uncorrelated with both `lemp` and `ldnpt`. This is likely violated since more productive firms may choose different input levels.

```
// Declare the panel structure: firm id 'index' and time variable 'yr'
xtset index yr

// Compute the Random Effects estimator for ldsal on lemp and ldnpt
xtreg ldsal lemp ldnpt, re

// Optionally, display robust standard errors
xtreg ldsal lemp ldnpt, re vce(robust)
```

Solution (d).
To eliminate the unobserved firm-specific effect α_i, we apply the within (time-demeaning) transformation. Define the time averages for each firm i as:

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} \text{ldsal}_{it}, \quad \bar{x}_{i1} = \frac{1}{T}\sum_{t=1}^{T} \text{lemp}_{it}, \quad \bar{x}_{i2} = \frac{1}{T}\sum_{t=1}^{T} \text{ldnpt}_{it},$$

and

$$\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}.$$

Subtracting these averages from the original model gives:

$$\text{ldsal}_{it} - \bar{y}_i = \beta_1\Big(\text{lemp}_{it} - \bar{x}_{i1}\Big) + \beta_2\Big(\text{ldnpt}_{it} - \bar{x}_{i2}\Big) + \Big(u_{it} - \bar{u}_i\Big).$$

Let the demeaned variables be:

$$\tilde{y}_{it} = \text{ldsal}_{it} - \bar{y}_i, \quad \tilde{\text{lemp}}_{it} = \text{lemp}_{it} - \bar{x}_{i1}, \quad \tilde{\text{ldnpt}}_{it} = \text{ldnpt}_{it} - \bar{x}_{i2},$$

and

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

Thus, the transformed model is:

$$\tilde{y}_{it} = \beta_1 \tilde{\text{lemp}}_{it} + \beta_2 \tilde{\text{ldnpt}}_{it} + \tilde{u}_{it}.$$

The FE within estimator is obtained by applying Ordinary Least Squares (OLS) to the demeaned model:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it} \right),$$

where $\tilde{x}_{it} = \begin{pmatrix} \tilde{\text{lemp}}_{it} \\ \tilde{\text{ldnpt}}_{it} \end{pmatrix}$.

Assumptions for Consistency

For the FE estimator to be consistent, the following assumptions must hold:

1. **Strict Exogeneity:**

$$E\Big(u_{it} \mid \text{lemp}_{i1}, \text{ldnpt}_{i1}, \ldots, \text{lemp}_{iT}, \text{ldnpt}_{iT}, \alpha_i\Big) = 0 \quad \text{for all } t.$$

2. **No Perfect Collinearity:** The demeaned regressors $\tilde{\text{lemp}}_{it}$ and $\tilde{\text{ldnpt}}_{it}$ are not perfectly collinear.

3. **Large Number of Cross-Sections:** Consistency is achieved as $N \to \infty$ while T remains fixed.

Under these conditions, the FE estimator is consistent even if the individual effects α_i are correlated with the regressors in the original model.

- The FE-W estimator is computed by demeaning the data to remove the firm-specific effect α_i.
- **Results (balanced panel):**
  - β_1: approximately 0.765.
  - β_2: approximately 0.408.
- **Interpretation:** Within a firm, a 1% increase in labor results in about a 0.765% increase in sales, and a 1% increase in capital results in about a 0.408% increase in sales. The sum (approximately 1.17) suggests increasing returns to scale in the within-firm variation.

Table 2: Question (d): Fixed Effects (Within) Estimator

	(1) FE		(2) FE Robust	
	b	se	b	se
log of employment	0.765	0.062	0.765	0.079
log of deflated capital	0.408	0.047	0.408	0.060
Observations	856		856	

- **Assumptions:** Requires strict exogeneity of the error term (i.e., no feedback from shocks to future inputs) after controlling for α_i. Although more realistic than RE, this condition may still be violated in practice.

```
xtreg ldsal lemp ldnpt , fe

// Optionally , use robust standard errors
xtreg ldsal lemp ldnpt , fe vce(robust)
```

Solution (e).

To eliminate the unobserved firm-specific effect α_i, we take first differences. For $t \geq 2$, define the first-differenced variables as:

$$\Delta \text{ldsal}_{it} = \text{ldsal}_{it} - \text{ldsal}_{i,t-1},$$

$$\Delta \text{lemp}_{it} = \text{lemp}_{it} - \text{lemp}_{i,t-1}, \quad \Delta \text{ldnpt}_{it} = \text{ldnpt}_{it} - \text{ldnpt}_{i,t-1},$$

$$\Delta u_{it} = u_{it} - u_{i,t-1}.$$

Then, the differenced model is given by:

$$\Delta \text{ldsal}_{it} = \beta_1 \, \Delta \text{lemp}_{it} + \beta_2 \, \Delta \text{ldnpt}_{it} + \Delta u_{it}.$$

Denote $\Delta y_{it} = \Delta \text{ldsal}_{it}$ and

$$\Delta x_{it} = \begin{pmatrix} \Delta \text{lemp}_{it} \\ \Delta \text{ldnpt}_{it} \end{pmatrix}.$$

The FE First Difference estimator is then:

$$\hat{\beta}_{FD} = \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta x_{it} \, \Delta x_{it}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta x_{it} \, \Delta y_{it} \right).$$

For $\hat{\beta}_{FD}$ to be consistent, the following assumptions are required:

1. **Strict Exogeneity in Differences:**

$$E\left(\Delta u_{it} \mid \Delta x_{i1}, \Delta x_{i2}, \dots, \Delta x_{iT}\right) = 0 \quad \text{for all } t.$$

2. **No Perfect Collinearity:** The differenced regressors must not be perfectly collinear.

3. **Large Number of Cross-Sections:** The estimator is consistent as $N \to \infty$ while T remains fixed.

- The FE-FD estimator is obtained by differencing consecutive observations, eliminating α_i.

- **Results (balanced panel):**
  - β_1: approximately 0.858.
  - β_2: approximately 0.184.

- **Interpretation:** Year-to-year changes indicate that a 1% increase in labor is associated with about a 0.86% increase in sales, while a 1% increase in capital is associated with only a 0.18% increase. The large difference in capital elasticity compared to the FE-W estimator may suggest that capital adjustments have a slower or more complex dynamic in the short-run.

- **Assumptions:** Consistency depends on the absence of serial correlation in the errors and that changes in inputs are not influenced by past shocks.

Solution (f).
Define the time-demeaned variables as:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \tilde{x}_{it} = x_{it} - \bar{x}_i,$$

with

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}.$$

Then the transformed model becomes:

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{u}_{it}, \quad \text{where} \quad \tilde{u}_{it} = u_{it} - \bar{u}_i.$$

The Fixed Effects (FE) estimator is given by:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{y}_{it}\right).$$

Since $\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{u}_{it}$, we have:

$$\hat{\beta}_{FE} - \beta = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it}\right).$$

Under standard regularity conditions, and with T fixed and $N \to \infty$, it holds that:

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}' \xrightarrow{p} Q,$$

and

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it} \xrightarrow{d} N(0, \Sigma),$$

where

$$Q = E\left[\frac{1}{T} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right],$$

and

$$\Sigma = \text{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it}\right).$$

Hence, by Slutsky's theorem,

$$\sqrt{N}(\hat{\beta}_{FE} - \beta) = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it}\right)$$

converges in distribution to:

$$\sqrt{N}(\hat{\beta}_{FE} - \beta) \xrightarrow{d} N\Big(0,\, Q^{-1}\Sigma Q^{-1}\Big).$$

Solution (g).
Recall from the asymptotic distribution of the FE estimator that

$$\sqrt{N}(\hat{\beta}_{FE} - \beta) \xrightarrow{d} N\Big(0,\, Q^{-1}\Sigma Q^{-1}\Big),$$

where

$$Q = \lim_{N\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}', \quad \Sigma = \lim_{N\to\infty} \text{Var}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it}\right).$$

An estimator for the asymptotic variance of $\hat{\beta}_{FE}$ is given by:

$$\widehat{\text{Var}}(\hat{\beta}_{FE}) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\hat{u}_{it}^2\right) \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1},$$

where $\hat{u}_{it}$ are the residuals from the FE regression. Consequently, the standard error for the jth coefficient is given by:

$$\text{se}(\hat{\beta}_j) = \sqrt{\left[\widehat{\text{Var}}(\hat{\beta}_{FE})\right]_{jj}}.$$

This robust variance estimator is used in empirical software to report standard errors that are consistent even under heteroscedasticity and autocorrelation of the error term.

- The robust (clustered) standard errors are computed using the formula

$$\hat{V} = \widehat{Q}^{-1}\widehat{\Sigma}\,\widehat{Q}^{-1},$$

  where $\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} X_i^{*\prime}\hat{u}_i\hat{u}_i' X_i^*$.

- With the balanced panel, the estimated standard errors are approximately:
  - SE($\hat{\beta}_1$) ≈ 0.0789,
  - SE($\hat{\beta}_2$) ≈ 0.0597.

Table 3: Question (g): FE Regression with Robust SE

	(1)	
	log of deflated sales	
	b	se
log of employment	0.765	0.079
log of deflated capital	0.408	0.060
Observations	856	

- These values indicate high statistical significance of the estimated coefficients.

```
xtreg ldsal lemp ldnpt , fe robust
```

Solution (h).
Consider a balanced panel with N firms and T time periods per firm. The Fixed Effects (FE) estimator is obtained by first demeaning the data:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \tilde{x}_{it} = x_{it} - \bar{x}_i,$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}.$$

The FE estimator is then given by

$$\hat{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}\right).$$

To account for within-firm correlation and obtain robust standard error estimates, we employ a clustered bootstrap procedure as follows:

1. **Resampling:** For $b = 1, \dots, B$ (with $B = 1000$), draw a bootstrap sample by resampling N firms with replacement from the original sample. For each selected firm, include all T observations.

2. **Estimation:** Compute the FE estimator on the bootstrap sample to obtain $\hat{\beta}^{*(b)}$.

3. **Variance Estimation:** The bootstrap estimator for the variance of $\hat{\beta}$ is

$$\widehat{\text{Var}}_{\text{boot}}(\hat{\beta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^{*(b)} - \bar{\hat{\beta}}^{*}\right)\left(\hat{\beta}^{*(b)} - \bar{\hat{\beta}}^{*}\right)',$$

where

$$\bar{\hat{\beta}}^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{*(b)}.$$

4. **Standard Errors:** The standard error for the jth coefficient is

$$\text{se}(\hat{\beta}_j) = \sqrt{\left[\widehat{\text{Var}}_{\text{boot}}(\hat{\beta})\right]_{jj}}.$$

This clustered bootstrap procedure yields standard error estimates that are robust to within-cluster (firm) correlation.

Table 4: Question (h): FE Bootstrap Estimates (Manual)

	(1)	
	b	se
_bs_1	0.765	0.081
_bs_2	0.408	0.058
Observations	856	

- The clustered bootstrap resamples firms (clusters) with replacement, preserving the panel structure.
- With $B = 1000$ bootstrap replications, the bootstrap standard errors are found to be:
  - $\text{SE}(\hat{\beta}_1) \approx 0.0812$,
  - $\text{SE}(\hat{\beta}_2) \approx 0.0582$.
- The bootstrap results confirm the accuracy of the analytical cluster-robust standard errors.

```
use "GMdata_balanced.dta", clear

program define fe_est, rclass
    preserve
    bysort index: egen ymean = mean(ldsal)
    bysort index: egen x1mean = mean(lemp)
    bysort index: egen x2mean = mean(ldnpt)

    gen yd = ldsal - ymean
    gen x1d = lemp - x1mean
    gen x2d = ldnpt - x2mean

    quietly regress yd x1d x2d, noconstant
    return scalar b1 = _b[x1d]
    return scalar b2 = _b[x2d]
    restore
end

bootstrap r(b1) r(b2), reps(1000) cluster(index) nodots: fe_est
```

Solution (i).

Consider the panel data model:

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}, \quad i = 1, \ldots, N,\ t \in \mathcal{T}_i,$$

where $\mathcal{T}_i$ denotes the set of time periods for which firm i is observed and $T_i = |\mathcal{T}_i|$. For each firm i, define the time averages based on the available observations:

$$\bar{y}_i = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i} y_{it}, \quad \bar{x}_i = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i} x_{it}.$$

The within (demeaning) transformation yields:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \tilde{x}_{it} = x_{it} - \bar{x}_i, \quad t \in \mathcal{T}_i.$$

The transformed model is:

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{u}_{it}, \quad \text{with} \quad \tilde{u}_{it} = u_{it} - \bar{u}_i,$$

and

$$\bar{u}_i = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i} u_{it}.$$

The Fixed Effects (FE) estimator is then given by:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N}\sum_{t\in\mathcal{T}_i} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t\in\mathcal{T}_i} \tilde{x}_{it}\tilde{y}_{it}\right).$$

Under standard assumptions (strict exogeneity, no perfect collinearity, etc.), $\hat{\beta}_{FE}$ is consistent for β even when the panel is unbalanced.

- The full dataset (unbalanced panel) is used to re-estimate the FE model.
- **Results (unbalanced panel):**

Table 5: Question (i): FE Estimator on Unbalanced Panel

	(1) FE		(2) FE Robust	
	b	se	b	se
log of employment	0.753	0.035	0.753	0.061
log of deflated capital	0.311	0.026	0.311	0.037
Observations	2971		2971	

  - $\beta_1 \approx 0.753$ with a clustered standard error of about 0.0607.
  - $\beta_2 \approx 0.311$ with a clustered standard error of about 0.0368.

- **Interpretation:** Compared to the balanced panel, the labor coefficient is similar, but the capital coefficient is lower. This may reflect sample selection effects; firms with incomplete data (often newer or smaller firms) exhibit lower capital productivity.

```
use "GMdata.dta", clear
xtset index yr

xtreg ldsal lemp ldnpt, fe

xtreg ldsal lemp ldnpt, fe vce(robust)
```

Solution (j).
Let $\hat{\beta}_{FE}$ be the Fixed Effects (FE) estimator obtained from the original panel data model:

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}, \quad i = 1, \ldots, N, \ t \in \mathcal{T}_i.$$

For each firm i, define the time-demeaned variables as:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \tilde{x}_{it} = x_{it} - \bar{x}_i,$$

with

$$\bar{y}_i = \frac{1}{T_i} \sum_{t \in \mathcal{T}_i} y_{it}, \quad \bar{x}_i = \frac{1}{T_i} \sum_{t \in \mathcal{T}_i} x_{it}.$$

The FE estimator is then given by:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N} \sum_{t \in \mathcal{T}_i} \tilde{x}_{it} \tilde{x}_{it}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t \in \mathcal{T}_i} \tilde{x}_{it} \tilde{y}_{it} \right).$$

To account for within-firm correlation, we use a clustered bootstrap procedure:

1. **Resampling:** For $b = 1, \ldots, B$ (with $B = 1000$), draw a bootstrap sample by sampling N firms with replacement from the original data. For each selected firm, include all T_i observations.

2. **Estimation:** Compute the FE estimator on the bth bootstrap sample to obtain $\hat{\beta}_{FE}^{*(b)}$.

3. **Bootstrap Variance:** The variance of $\hat{\beta}_{FE}$ is estimated by

$$\widehat{\text{Var}}_{\text{boot}}(\hat{\beta}_{FE}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_{FE}^{*(b)} - \bar{\hat{\beta}}^* \right) \left(\hat{\beta}_{FE}^{*(b)} - \bar{\hat{\beta}}^* \right)',$$

   where

$$\bar{\hat{\beta}}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_{FE}^{*(b)}.$$

4. **Standard Errors:** The standard error for the jth coefficient is given by

$$\text{se}(\hat{\beta}_j) = \sqrt{\left[\widehat{\text{Var}}_{\text{boot}}(\hat{\beta}_{FE}) \right]_{jj}}.$$

This clustered bootstrap method provides standard error estimates that are robust to intra-cluster (firm-level) correlation.

Table 6: Question (j): FE Bootstrap on Unbalanced Panel (Manual)

	(1)	
	b	se
_bs_1	0.753	0.061
_bs_2	0.311	0.038
Observations	2971	

- The clustered bootstrap is also applied to the unbalanced panel.
- With $B = 1000$ replications, the bootstrapped standard errors are approximately:
  - $\text{SE}(\hat{\beta}_1) \approx 0.0609$,
  - $\text{SE}(\hat{\beta}_2) \approx 0.0379$.

- These values closely match the analytical robust standard errors, reinforcing the robustness of our estimates.

```
use "GMdata.dta", clear
xtset index yr

program define fe_est2, rclass
    preserve
    bysort index: egen ymean2 = mean(ldsal)
    bysort index: egen x1mean2 = mean(lemp)
    bysort index: egen x2mean2 = mean(ldnpt)

    gen yd2 = ldsal - ymean2
    gen x1d2 = lemp - x1mean2
    gen x2d2 = ldnpt - x2mean2

    quietly regress yd2 x1d2 x2d2, noconstant
    return scalar b1 = _b[x1d2]
    return scalar b2 = _b[x2d2]
    restore
end

bootstrap r(b1) r(b2), reps(1000) cluster(index) nodots: fe_est2
```