Financial Econometrics
Shanghai Jiao Tong University
Summer School – 2026
Empirical Exercise 10.2
With Solutions

Fatih Kansoy

Reference: Stock, J.H. and Watson, M.W., Introduction to Econometrics (4th Edition, Global Edition)

Data: `income_democracy.csv`*, a panel of countries observed in 1960, 1965, . . . , 2000. The democracy index* `Dem_ind` *is the column* `dem_ind`*,* `Log_GDPPC` *is the column* `log_gdppc`*, and the country identifier is the column* `code`*.*

Empirical Exercise 10.2. Do citizens demand more democracy and political freedom as their incomes grow? That is, is democracy a normal good? In this exercise you will use the data file `income_democracy.csv`, a panel data set covering countries for the years 1960, 1965, . . . , 2000. The data set contains an index of political freedom/democracy for each country in each year, together with data on the country's income and various demographic controls. (The income and demographic controls are lagged five years relative to the democracy index to allow time for democracy to adjust to changes in these variables.) The democracy index is the column `dem_ind` (S&W's `Dem_ind`), the logarithm of per capita income is the column `log_gdppc` (S&W's `Log_GDPPC`), and the country identifier is the column `code`.

(a) Is the data set a balanced panel? Explain.

(b) The index of political freedom/democracy is labelled `Dem_ind`.

(i) What are the minimum and maximum values of `Dem_ind` in the data set? What are the mean and standard deviation of `Dem_ind` in the data set? What are the 10th, 25th, 50th, 75th, and 90th percentiles of its distribution?

(ii) What is the value of `Dem_ind` for the United States in 2000? Averaged over all years in the data set?

(iii) What is the value of `Dem_ind` for Libya in 2000? Averaged over all years in the data set?

(iv) List five countries with an average value of `Dem_ind` greater than 0.95; less than 0.10; and between 0.3 and 0.7.

(c) The logarithm of per capita income is labelled `Log_GDPPC`. Regress `Dem_ind` on `Log_GDPPC`. Use standard errors that are clustered by country.

(i) How large is the estimated coefficient on `Log_GDPPC`? Is the coefficient statistically significant?

(ii) If per capita income in a country increases by 20%, by how much is `Dem_ind` predicted to increase? What is a 95% confidence interval for the prediction? Is the predicted increase in `Dem_ind` large or small? (Explain what you mean by large or small.)

(iii) Why is it important to use clustered standard errors for the regression? Do the results change if you do not use clustered standard errors?

(d)

(i) Suggest a variable that varies across countries but plausibly varies little, or not at all, over time and that could cause omitted variable bias in the regression in (c).

(ii) Estimate the regression in (c), allowing for country fixed effects. How do your answers to (c)(i) and (c)(ii) change?

(iii) Exclude the data for Azerbaijan and rerun the regression. Do the results change? Why or why not?

(iv) Suggest a variable that varies over time but plausibly varies little, or not at all, across countries and that could cause omitted variable bias in the regression in (c).

(v) Estimate the regression in (c), allowing for time and country fixed effects. How do your answers to (c)(i) and (c)(ii) change?

(vi) There are additional demographic controls in the data set. Should these variables be included in the regression? If so, how do the results change when they are included?

(e) Based on your analysis, what conclusions do you draw about the effects of income on democracy?

Solution:

Data Overview

The file `income_democracy.csv` is a country-year panel covering the years 1960, 1965, ..., 2000. The key variables are:

Variable	Column	Description
`Dem_ind`	`dem_ind`	Index of political freedom/democracy (0 to 1)
`Log_GDPPC`	`log_gdppc`	Logarithm of per capita income (lagged 5 years)
Country identifier	`code`	Used as the entity for clustering and fixed effects
Controls	various	Age, education, and population demographic variables

Note: The income and demographic controls are lagged five years relative to the democracy index. This gives democracy time to respond to changes in income and demographics. Throughout, the country (`code`) is the entity used both for clustering the standard errors and, where stated, for country fixed effects.

The five regressions referenced below are summarised here. The dependent variable is `Dem_ind`; the reported figure on `log_gdppc` is the coefficient, with the clustered standard error in parentheses and the 95% confidence interval in brackets.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
`log_gdppc`	0.236	0.235	0.083	0.054	0.025
	(0.012)	(0.012)	(0.031)	(0.042)	(0.054)
	[0.212,	[0.211,	[0.021,	[−0.030,	[−0.057,
	0.259]	0.259]	0.146]	0.137]	0.120]
Controls	No	No	No	No	Yes
Country effects	No	No	Yes	Yes	Yes
Time effects	No	Yes	No	Yes	Yes

The controls in column (5) are the age, education, and population variables. The relevant exclusion tests are: time effects $F = 9.31$ in (2), 5.73 in (4) and 4.61 in (5) (all $p < 0.001$); the age variables $F = 2.12$ ($p = 0.08$) in (5); and the joint age, education, and population variables $F = 1.44$ ($p = 0.21$) in (5).

(a) Is the data set a balanced panel?

No, the data set is not a balanced panel. A balanced panel would have every country observed in every year (1960, 1965, . . . , 2000), giving the same nine observations per country. Here data are available over different years for different countries, so it is an **unbalanced panel**.

Illustration: `Dem_ind` for Andorra is available only for 1970, 1995, and 2000, whereas for Afghanistan it is available for all nine years (1960, 1965, . . . , 2000). Because the time coverage differs across countries, the panel is unbalanced.

The fixed effects methods of this chapter still apply to an unbalanced panel; the entity-demeaning is simply carried out over whatever years are available for each country.

(b) Describing the democracy index `Dem_ind`

(i) Distribution of `Dem_ind`

Statistic	Value
Minimum	0.00
Maximum	1.00
Mean	0.50
Standard deviation	0.37

Percentile	Value
10th	0.00
25th	0.17
50th (median)	0.50
75th	0.83
90th	1.00

The index runs over the full $[0, 1]$ range. The large standard deviation (0.37) relative to the mean (0.50), together with 10th and 90th percentiles sitting exactly at the bounds 0 and 1, tells you that the distribution is very spread out: many country-years are at, or close to, full autocracy (0) or full democracy (1).

(ii) United States

For the United States in 2000, `Dem_ind` = 1.00. Averaged over the nine years in the sample, the value is **0.99**. The United States is therefore at, or essentially at, the top of the index throughout.

(iii) Libya

For Libya in 2000, `Dem_ind` = 0.00. Averaged over the nine years in the sample, the value is **0.11**. Libya sits near the bottom of the index throughout the period.

(iv) Listing countries by average `Dem_ind`

This part asks you to read through the data and group countries by their country average of `Dem_ind`: five with an average greater than 0.95 (stable democracies), five with an average below 0.10 (persistent autocracies), and five with an average between 0.3 and 0.7 (mixed or transitional regimes). Compute the within-country mean of `dem_ind` grouped by `code`, sort, and pick examples from each band. As the official key puts it, this is an instructive exercise to do directly with the data; the precise five names are not unique.

(c) Regression of `Dem_ind` on `Log_GDPPC` (clustered by country)

This is column (1) of the summary table. With standard errors clustered by `code`:

$$\boxed{\widehat{\texttt{Dem\_ind}} = \hat{\beta}_0 + 0.236 \times \texttt{Log\_GDPPC}}$$

(i) Size and significance of the coefficient

Regressor	Coefficient	Clustered S.E.	95% CI	Significant?
`Log_GDPPC`	0.236	0.012	[0.212, 0.259]	Yes (1% level)

The estimated coefficient is 0.236 (about 0.24) with a clustered standard error of 0.012 (about 0.01). The 95% confidence interval is roughly 0.21 to 0.26. Because $t = 0.236/0.012 \approx 20$, the coefficient is very strongly statistically significant. Taken at face value it is also a large effect, as explained in (ii).

(ii) Predicted effect of a 20% rise in income

A 20% increase in per capita income raises log(GDPPC) by approximately $\ln(1.20) \approx 0.20$. The predicted change in the democracy index is therefore:

$$\Delta\widehat{\texttt{Dem\_ind}} \approx 0.20 \times 0.236 = \mathbf{0.048}.$$

Using the endpoints of the confidence interval for the slope, a 95% confidence interval for this predicted change is approximately:

$$0.20 \times 0.21 \text{ to } 0.20 \times 0.26 \;=\; [\,0.042,\; 0.052\,].$$

Is this large or small? A natural yardstick is the standard deviation of `Dem_ind` in the data, which is 0.37. The predicted change 0.048 is about $0.048/0.37 \approx 1/10$ of one standard deviation. By that benchmark, a 20% income rise moves a country roughly one-tenth of a standard deviation up the democracy index, which is a **large** effect: a sustained run of income growth would, on this estimate, move countries appreciably along the democracy scale.

(iii) Why clustered standard errors?

Clustered standard errors are needed because of **country-specific omitted factors**. The same country is observed in multiple years, and the regression errors are correlated over time within a country (positive within-country, or serial, correlation). Clustering by `code` allows for arbitrary correlation of the errors within a country while treating different countries as independent.

The results do change if you ignore clustering. The **clustered** standard error on `Log_GDPPC` is 0.012; the ordinary (unclustered) standard error is smaller, about 0.007, because it incorrectly ignores the positive within-country correlation of the errors. Using the unclustered standard error would overstate the precision of the estimate and understate the width of the confidence interval.

(d) Omitted variables and fixed effects

(i) A variable that varies across countries but not over time

Countries differ in their **histories, institutions, social structures, and religions**. These features change slowly, if at all, over the sample period, yet they plausibly affect a country's preference for democracy and may also be correlated with economic development (and hence per capita income). Any such variable, if omitted, can cause omitted variable bias in the simple regression of (c). This motivates including **country fixed effects**.

(ii) Adding country fixed effects

This is column (3) of the summary table.

$$\boxed{\texttt{Dem\_ind}_{it} = \beta_1 \, \texttt{Log\_GDPPC}_{it} + \alpha_i + u_{it}}$$

	Coefficient	Clustered S.E.	95% CI	Significant?
Log_GDPPC	0.083	0.032	[0.021, 0.146]	Yes (1% level)

The estimated coefficient **falls by a factor of about three**, from 0.236 to 0.083, with a standard error of 0.032. Although the estimated effect is much smaller, it remains statistically significant at the 1% level. The implied answer to (c)(ii) also falls by the same factor of three: a 20% income rise is now predicted to raise Dem_ind by about $0.20 \times 0.083 \approx 0.017$, roughly one-third of the earlier estimate. The sharp drop suggests that much of the cross-country association between income and democracy in (c) reflects fixed country characteristics rather than a within-country effect of income.

(iii) Excluding Azerbaijan

Data for Azerbaijan are available **only for the year 2000**. In a country fixed effects regression, a country observed in a single year is **completely absorbed by its own fixed effect**: the single observation is exactly fitted by the country intercept and contributes nothing to the within-country variation used to estimate β_1. Therefore dropping Azerbaijan **does not change the estimated coefficient on** log_gdppc. (More generally, any country with only one observation is irrelevant to the fixed effects slope.)

(iv) A variable that varies over time but not across countries

The **demand for democracy can be contagious and sweep across countries** at a point in time. A prominent example is the wave of pro-democracy movements often referred to as the "Arab Spring". Such global waves are common across countries in

a given period but vary over time. To the extent that these episodes coincide with global economic conditions, they are correlated with `log_gdppc` and so can cause omitted variable bias. This motivates including **time fixed effects**.

(v) Adding both time and country fixed effects

This is column (4) of the summary table.

$$\boxed{\texttt{Dem\_ind}_{it} = \beta_1 \texttt{Log\_GDPPC}_{it} + \alpha_i + \lambda_t + u_{it}}$$

	Coefficient	Clustered S.E.	95% CI	Significant?
`Log_GDPPC`	0.054	0.042	$[-0.030, 0.137]$	No (not at 10%)

The estimated coefficient **falls further to** 0.054, approximately 1/5 of the value from the regression that omits both time and country fixed effects. The standard error is 0.042, the confidence interval $[-0.030, 0.137]$ now includes zero, and the estimate is **not statistically significant at the 10% level**. The time effects are jointly significant ($F = 5.73$, $p < 0.001$), confirming that common time shocks matter. The implied prediction for (c)(ii) shrinks to about $0.20 \times 0.054 \approx 0.011$, with a confidence interval that includes zero.

(vi) Adding the demographic controls

This is column (5) of the summary table, which adds the population, age, and education controls.

	Coefficient	Clustered S.E.	95% CI	Significant?
`Log_GDPPC`	0.025	0.054	$[-0.057, 0.120]$	No (not at 10%)

Jointly, the demographic variables are **not statistically significant**: the F-statistic for the age, education, and population variables together is 1.44 ($p = 0.21$), though the age variables alone are significant at the 10% level ($F = 2.12$, $p = 0.08$). When these controls are included, the estimated coefficient on `log_gdppc` falls further to 0.025 with a standard error of 0.054, so it is again not significant. Because the demographic block is jointly insignificant, there is no strong statistical case for retaining it; its main effect here is to confirm that the income coefficient is small and imprecisely estimated once fixed effects are present.

(e) Conclusions: does income raise democracy?

The headline pattern is a steady erosion of the income effect as more sources of omitted variable bias are controlled for:

Specification	**Coef.**	**Clustered S.E.**	**Significant?**	**Column**
Pooled (no effects)	0.236	0.012	Yes (1%)	(1)
Country fixed effects	0.083	0.032	Yes (1%)	(3)
Country + time effects	0.054	0.042	No	(4)
+ demographic controls	0.025	0.054	No	(5)

Conclusion: After controlling for omitted variables, and in particular for country fixed effects (and then time fixed effects and demographic controls), there is **little evidence of an income effect on the demand for democracy**. The large, highly significant association in the pooled regression largely reflects fixed differences across countries, not a within-country response of democracy to income.

Summary: Key Takeaways

Question	Finding	Interpretation
Balanced panel?	No	Different years for different countries
Range of `Dem_ind`	0 to 1, mean 0.50, s.d. 0.37	Very dispersed across regimes
Pooled income coef.	0.236 (s.e. 0.012)	Large and highly significant
+ country effects	0.083 (s.e. 0.032)	Falls by $\approx 3\times$, still significant
+ time effects	0.054 (s.e. 0.042)	Not significant; CI includes 0
+ demographics	0.025 (s.e. 0.054)	Not significant; controls jointly weak
Azerbaijan dropped	No change	Single-year country absorbed by FE
Overall	Weak income effect	Cross-country gap $\neq$ within-country effect

The Big Picture

This exercise illustrates several important panel data ideas:

1. **Unbalanced panels are routine.** Fixed effects methods still apply; entity-demeaning simply uses whatever years each country supplies.

2. **Fixed effects strip out level differences.** Country fixed effects absorb time-invariant features (history, institutions, religion). The fall from 0.236 to 0.083 shows how much of the raw association was driven by such fixed differences rather than by income changing democracy within a country.

3. **Time effects absorb common shocks.** Global waves of democratisation that coincide with world economic conditions are a credible source of bias; adding time effects pushes the income coefficient below conventional significance.

4. **Single-observation entities are uninformative for the within estimator.** Azerbaijan, observed only in 2000, is perfectly fitted by its own fixed effect and so cannot move the slope.

5. **Clustered standard errors matter.** Ignoring within-country serial correlation understates the standard error (0.007 versus 0.012), giving a false sense of precision.

6. **Identification drives the conclusion.** A causal-looking pooled result can evaporate once you ask whether the variation used to identify it is within country or merely across countries.

Calculations performed using Stata. See accompanying `E10_2_solution.do` *file.*