POLS 607 — Instructor: Casey Crisman-Cox
Panel Data — Spring 2025

Problem Set 2

This problem set is due February 27th. Upload your write up and any code files to the Github Classroom before midnight. You may work together, but you must turn in separately written (unique) write ups and/or code.

In this problem set, we are going to consider the relationship between economic growth and civil war. We will base on study on data and analysis from Miguel, Satynanath, and Sergenti (2004). A copy of this paper is included along with a code book describing their data. The dataset is found in the file `conflict.dta`.

1. Let's start on the theoretical end of this relationship and consider the following model for the relationship between civil conflict w_{it} and economic growth g_{it}

$$w_{it} = \alpha_i + \tau_t + \beta_1 g_{it} + \beta_2' x_{it} + \varepsilon_{it} \tag{1}$$
$$g_{it} = \kappa_i + \rho_t + \gamma_1 w_{it} + \gamma_2' x_{it} + u_{it}, \tag{2}$$

where x_{it} is some length-k vector of control variables. Additionally, let's add some structure in the sense of imposing weak exogeneity

$$\mathrm{E}[\varepsilon_{it} x_{it}] = \mathrm{E}[u_{it} x_{it}] = 0$$

and assuming that each unit i is independent of the others. For the within-estimator to be a consistent estimator of β_1 in Eq. 1, we will need to establish that $\mathrm{E}[\varepsilon_{it} g_{it}] = 0$.

(a) Rewrite equations 1 & 2 to be in reduced form. This means that w_{it} and g_{it} should be functions of only the exogenous variables x_{it}, composite parameters made up of combinations of the above parameters, and a composite error term. Another way to say this is to solve the pair of equations for w_{it} and g_{it} such that they only appear on the left-hand side.

The reduced form equations can be written as

$$w_{it} = \left(\frac{\alpha_i + \beta_1 \kappa_i}{1 - \beta_1 \gamma_1}\right) + \left(\frac{\tau_t + \beta_1 \rho_t}{1 - \beta_1 \gamma_1}\right) + \left(\frac{\beta_1 \gamma_2 + \beta_2}{1 - \beta_1 \gamma_1}\right)' x_{it} + \frac{\beta_1 u_{it} + \epsilon_{it}}{1 - \beta_1 \gamma_1}$$
$$g_{it} = \left(\frac{\kappa_i + \gamma_1 \alpha_i}{1 - \beta_1 \gamma_1}\right) + \left(\frac{\rho_t + \gamma_1 \tau_t}{1 - \beta_1 \gamma_1}\right) + \left(\frac{\gamma_1 \beta_2 + \gamma_2}{1 - \beta_1 \gamma_1}\right)' x_{it} + \frac{u_{it} + \gamma_1 \varepsilon_{it}}{1 - \beta_1 \gamma_1}.$$

(b) Find the expected value of $\mathrm{E}[\varepsilon_{it} g_{it}]$. What additional assumptions would be required to make it 0? Are any of them believable or plausible? Can the within estimator provide a consistent estimate β_1 if fit to Eq. 1?

For ease, suppose that the unit and time heterogeneity are fixed, then we get:

$$\begin{aligned}
\mathrm{E}[\varepsilon_{it} g_{it}] &= \frac{\mathrm{E}[\varepsilon_{it}](\kappa_i + \alpha_i \gamma_1)}{1 - \beta_1 \gamma_1} + \frac{\mathrm{E}[\varepsilon_{it}](\rho_t + \gamma_1 \tau_t)}{1 - \beta_1 \gamma_1} + \frac{\mathrm{E}[\varepsilon_{it} x_{it}]'(\beta_2 \gamma_1 + \gamma_2)}{1 - \beta_1 \gamma_1} \\
&\quad + \frac{\mathrm{E}[\varepsilon_{it}^2]\gamma_1}{1 - \beta_1 \gamma_1} + \frac{\mathrm{E}[\varepsilon_{it} u_{it}]}{1 - \beta_1 \gamma_1} \\
&= \frac{\mathrm{E}[\varepsilon_{it}^2]\gamma_1 + \mathrm{E}[\varepsilon_{it} u_{it}]}{1 - \beta_1 \gamma_1}.
\end{aligned}$$

For this to be 0, we would one of the following conditions:

i. $\gamma_1 = 0$ and $\mathrm{E}[\varepsilon_{it} u_{it}]$ or
ii. $\mathrm{E}[\varepsilon_{it} u_{it}] = -\mathrm{E}[\varepsilon_{it}^2]$.

The former is not particularly plausible given that we suspect there is a relationship between war and growth. The latter is a very particular covariance structure that we have no reason to suspect. As such, we would not expect the within estimator to be consistent for β_1.

(c) Suppose you fit the reduced form equations where you regress w_{it} on just x_{it} and g_{it} on just x_{it}. Could you use these regression coefficients to solve for $\hat{\beta}_1$? Why or why not?

The reduced form coefficients on x_{it} are length k so we have $2k$ equations to work with but there are $2k + 2$ unknowns. We will be able to solve them without additional structure.

(d) Let's separate x_{it} and it's original parameters as follows

$$\begin{aligned}
x_{it}' &= (z_{it}, q_{it}')' \\
\beta_2' &= (\phi_0, \phi_1')' \\
\gamma_2' &= (\psi_0, \psi_1')' \\
\beta_2' x_{it} &= z_{it}\phi_0 + \phi_1' q_{it} \\
\gamma_2' x_{it} &= z_{it}\psi_0 + \psi_1' q_{it},
\end{aligned}$$

where z_{it}, ϕ_0 and ψ_0 are scalars and q_{it}, ϕ_1 and ψ_1 are length-$k-1$ vectors. Rewrite your reduced form equations with these substitutions.

Making the substitutions, we get

$$w_i = \left(\frac{\alpha_i + \beta_1\kappa_i}{1-\beta_1\gamma_1}\right) + \left(\frac{\beta_1\psi_0 + \phi_0}{1-\beta_1\gamma_1}\right) z_{it} + \left(\frac{\beta_1\psi_1 + \phi_1}{1-\beta_1\gamma_1}\right)' q_{it} + \frac{\beta_1 u_{it} + \epsilon_{it}}{1-\beta_1\gamma_1}$$
$$g_i = \left(\frac{\kappa_i + \gamma_1\alpha_i}{1-\beta_1\gamma_1}\right) + \left(\frac{\gamma_1\phi_0 + \psi_0}{1-\beta_1\gamma_1}\right) z_{it} + \left(\frac{\gamma_1\phi_1 + \psi_1}{1-\beta_1\gamma_1}\right)' q_{it} + \frac{u_{it} + \gamma_1\varepsilon_{it}}{1-\beta_1\gamma_1}.$$

(e) Let $\phi_0 = 0$. Can you consistently estimate β_1 using a function of the reduced form estimates now? Why or why not? (**HINT:** The answer is yes, but you have to explain how). What does this tell you about what empirical moments identify $\hat{\beta}_1$? (i.e., what characteristics of the data actually determine $\hat{\beta}_1$?

Making this substitution we get

$$w_i = \left(\frac{\alpha_i + \beta_1\kappa_i}{1-\beta_1\gamma_1}\right) + \left(\frac{\beta_1\psi_0}{1-\beta_1\gamma_1}\right) z_{it} + \left(\frac{\beta_1\psi_1 + \phi_1}{1-\beta_1\gamma_1}\right)' q_{it} + \frac{\beta_1 u_{it} + \epsilon_{it}}{1-\beta_1\gamma_1}$$
$$g_i = \left(\frac{\kappa_i + \gamma_1\alpha_i}{1-\beta_1\gamma_1}\right) + \left(\frac{\psi_0}{1-\beta_1\gamma_1}\right) z_{it} + \left(\frac{\gamma_1\phi_1 + \psi_1}{1-\beta_1\gamma_1}\right)' q_{it} + \frac{u_{it} + \gamma_1\varepsilon_{it}}{1-\beta_1\gamma_1}.$$

We can now see that the ratio of coefficients for z_{it} gives us

$$\left(\frac{\beta_1\psi_0}{1-\beta_1\gamma_1}\right) \Big/ \left(\frac{\psi_0}{1-\beta_1\gamma_1}\right) = \beta_1.$$

This means that we can estimate β_1 using the ratio of these reduced-form estimates.

What we can see here is that β_1 can be estimated using the marginal effects of z on both w and g. This tells us that the covariances of z with w and g are what identify β_1. This is where LATE interpretations come into play as β_1 is only identified using information from countries where these covariances exist.

(f) Note that when $\phi_0 = 0$, z_{it} is an instrument for g_{it}, explain why it is valid and relevant.

It is relevant as it affects g_{it} ($\psi_0 \neq 0$) and it is valid because it does not have any effect on w_i either directly ($\phi_0 = 0$) or through x_{it} or ε_{it} (by assumption).

(g) Based on all of your insight above, we decide to fit the reduced-form system equations:

$$y = \mathbf{X}\begin{bmatrix}\lambda\\ \delta\end{bmatrix} + D_N\begin{bmatrix}c\\ k\end{bmatrix} + D_T\begin{bmatrix}s\\ r\end{bmatrix} + e,$$

where

$$y = \begin{bmatrix}w\\ g\end{bmatrix} \qquad D_N = \begin{bmatrix}I_N \otimes 1_T & 0\\ 0 & I_N \otimes 1_T\end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix}X & 0\\ 0 & X\end{bmatrix} \qquad D_T = \begin{bmatrix}1_T \otimes I_N & 0\\ 0 & 1_T \otimes I_N\end{bmatrix}$$

Here, $c = (c_i)_1^N$, $k = (k_i)_1^N$, $r = (r_t)_1^T$, $s = (s_t)_1^T$, λ, and δ are the reduced-form parameters that are composed of the parts you described above. Likewise, e is combined reduced-form error term, which you also described above. Let $\hat{V}$ be an appropriate estimated covariance matrix of $(\hat{\lambda}, \hat{\delta})$, find analytic expressions for $\hat{\beta}_1$ and its standard error from the reduced-form estimates found by fitting the system model. (**HINT:** Use the delta method to find the the standard error)

We have that $\hat{\beta}_1 = \hat{\lambda}_1/\hat{\delta}_1$ and so the variance becomes

$$\begin{aligned}\widehat{\text{Var}}(\hat{\beta}_1) &= (D_\theta \hat{\beta}_1)'\hat{V}(D_\theta \hat{\beta}_1)\\ &= (1/\hat{\delta}_1, 0, 0, -\hat{\lambda}_1/\hat{\delta}_1^2, 0, 0)\hat{V}(1/\hat{\delta}_1, 0, 0, -\hat{\lambda}_1/\hat{\delta}_1^2, 0, 0)'\end{aligned}$$

2. Let's turn to the data. For this problem you may use any packages you and pre-canned routines you want except where noted. I believe the main variables they use are as follows:

- `any_prio`: Is there an active civil conflict in country i, year t (0/1)? This comes from the PRIO database. An active conflict is one with at least 25 battle deaths per year
- `gdp_g`: Annual economic growth rate in country i, year t
- `gdp_g_l`: Annual economic growth rate in country i, year $t-1$
- `y_0`: Logged GDP per capita in country i in 1979
- `polity2l`: Polity 2 score in country i, year $t-1$

- `ethfrac`: Ethnolingustic fractionalization. Probability that two random citizens from country i are from the different ethnic groups. Based on Soviet ethnographic data.
- `relfrac`: Religious fractionalization. Probability that two random citizens from country i are from the different religious groups. Based on CIA factbook.
- `Oil`: Is more than a 1/3 of country i's export revenue from oil?
- `lmtnest`: The logged proportion of the country that is covered in mountains
- `lpopl1`: Logged population of country i at time $t-1$

(a) Using the above list, verify that that we've identified the right measures by replicating the following parts of Table 1

- The first row of section A
- Section C
- All but the last row of section D

The summary statistics are presented in Table 1.

Table 1: Summary statistics

	Mean	St. Dev.	Obs.
PRIO conflict	0.268	0.443	743
Growth	-0.005	0.071	743
Lagged growth	-0.006	0.072	743
Log(GDP p.c.), 1979	1.164	0.901	743
Democracy (lag)	-3.608	5.554	743
Eth. Frac.	0.655	0.237	743
Rel. Frac.	0.487	0.186	743
Oil exporter	0.118	0.323	743
Log(Mountainous terrain)	1.578	1.433	743
Log(Population) (lag)	8.75	1.207	743

(b) The authors are interested in the effect of economic growth on conflict. Do you think a specification where we regress `any_prio` on the other variables listed makes sense? Why or why not?

Yes, in the sense that it's what we're interested in and all of the other variables are reasonable controls. No, in the sense that we're fairly sure it's unidentified given that growth and civil conflict are certainly endogenous, but I suspect that with lags and fixed effects we'll get something not unreasonable as an association.

(c) Reproduce columns 2-4 of Table 4 as best you can. For each of these interpret the effect of growth at time t on conflict. Which model specification do you think is the most convincing? Why or why not?

The trick here is using years since 1978 as your time trend. Otherwise, you can get weird identification in the country-specific time trend model because it's using the overall intercept as the same intercept for each trend. It's weird that it makes a difference and it's weird that they still went with it. The results are presented in Table 2.

Note that growth in measured in 100ths of a percentage point and conflict is binary. In the first model, we see that, on average, a 1 percentage point increase in growth decreases the probability of conflict by 0.0033 (0.33 percentage points), holding the controls fixed. In latter two, we see that the immediate effect of 1 percentage point increase in growth is a decrease in the probability of conflict by about 0.21 percentage points, on average and hold all else equal. The fixed effects model is the most credible as it controls for a range of potentially unobserved factors.

(d) Refit the model from column 4 of Table 4 with two-way fixed effects instead of the country-specific time. Compare the standard errors from clustering on country, clustering on time, Driscoll-Kraay, and two-way clustering. What differences do you observe? Which you use if this was your paper. Justify your answer.

The standard error comparisons for the coefficients on growth are presented in Table 3. Here we see that clustering on country and two-way clustering produce the most conservative estimates. Clustering by year produces much smaller standard errors and Dricoll-Kraay splits the difference. Over all I would stick with just clustering on country because:

i. It is the level of treatment assignment

Table 2: Replicating the OLS models from Table 4

	(2)	(3)	(4)
Growth	−0.33	−0.21	−0.21
	(0.26)	(0.20)	(0.16)
Lagged growth	−0.08	0.01	0.07
	(0.24)	(0.20)	(0.16)
Log(GDP p.c.), 1979	−0.04	0.09	
	(0.05)	(0.08)	
Democracy	0.00	0.00	
	(0.00)	(0.01)	
Eth. Frac.	0.23	0.51	
	(0.27)	(0.40)	
Rel. Frac.	−0.24	0.10	
	(0.24)	(0.42)	
Oil exporter	0.05	−0.16	
	(0.21)	(0.20)	
Log(Mountainous terrain)	0.08	0.06	
	(0.04)	(0.06)	
Log(Population)	0.07	0.18	
	(0.05)	(0.09)	
Country fixed effects	no	no	yes
Country-specific time trend	no	yes	yes
R2	0.126	0.530	0.707
RMSE	0.41	0.30	0.24
Num.Obs.	743	743	743

ii. They're conservative

iii. Our N dimension is larger than our T (41 versus 19). Only 41 countries does not give us a lot to average over, but more than then 19 years do.

Table 3: Comparing standard errors

	Growth	Lagged growth
Estimate	-0.41	-0.19
Clustered by country	0.21	0.22
Clustered by year	0.18	0.17
Driscoll-Kraay	0.22	0.17
2-way clustering	0.22	0.22

(e) The authors propose using rainfall as an instrument for growth with the variables `GPCP_g` and `GPCP_g_l` measuring the average growth in rainfall in country i at times t and $t-1$, respectively. Briefly describe what we need to assume for these instruments to be valid and relevant for the two growth measures. Do you think these instruments are valid? Why or why not?

To be relevant, rainfall must correlate with growth. This is plausible although even in heavily agrarian economies this is likely to be weak. To be valid, rainfall should affect civil conflict only through growth. I'm skeptical of validity here because I suspect that heavy rain can also make battlefield terrain favor rebels and might affect conflict that way. Likewise, others have found that rain can directly affect decisions to riot or protest in ways that clearly outside of its economic impact (e.g., Sarsons 2015, *Journal of Development Economics*).

(f) Use 2SLS to fit a model where we regress `any_prio` on `gdp_g` using `GPCP_g` as an instrument. Control for democracy and population and use two-way fixed effects. Verify your analysis from question 1 by then using a system estimator and delta method to get to the same estimate for the effect of growth on conflict.

NOTE: We want to make sure that we have different fixed effects for the equations, *but* we don't want the standard errors to be over confident by double counting the number of units. An easy way to avoid problems is to create new country and year variables that are equal to `ccode` and `year` for the part of the data that is used in equation 1, and equal to that plus 500 (or some other large enough number) otherwise. Then cluster on the original `ccode` variable (or `year` or both). If you've done this right the estimates and standard errors should match the results of fitting the equations separately (subject to small differences in degrees of free in the standard errors).

Table 4: Identifying β_1

	Est	St. Err.
2SLS	1.79	1.35
System	1.79	1.33

(g) Having satisfied yourself that we understand the identification of this approach, go ahead and attempt to reproduce columns 5-6 from Table 4. Interpret the effect of growth on conflict.

The results are presented in Table 5. For the contemporaneous effect of growth we see that in the first (second) model, a 1 percentage point increase in growth decreases the probability of conflict by about 0.41 (1.13) percentage points, on average and holding the other variables constant.

(h) Consider the first stage regressions of rainfall on growth. Are these strong instruments?

The first stage F statistics for these two models are presented in Table 6. Overall, these appear to be relatively weak instruments.

(i) Consider an overidentified model where `GPCP_g` and `GPCP_g_l` instrument for just `GPCP_g` (i.e., drop `GPCP_g_l`). Fit this model with time-varying controls and two-way fixed effects. Fit the model with GMM using a weighting matrix and covariance matrix that you think is appropriate, justify your answer. Interpret the results and test the hypothesis that instruments are valid.

The overidentified GMM results are presented in Table 7. Here we see that a one percentage point increase contemporaneous growth is associated with a 0.26 percentage point increase in conflict risk, on average, although the estimate is very imprecise. The Sargan test suggests that these instruments are not valid for contemporaneous growth.

Table 5: Replicating 2SLS results from MSS

	(5)	(6)
Growth	−0.41	−1.13
	(1.48)	(1.36)
Lagged growth	−2.25	−2.55
	(1.07)	(1.07)
Log(GDP p.c.), 1979	0.05	
	(0.10)	
Democracy	0.00	
	(0.01)	
Eth. Frac.	0.51	
	(0.39)	
Rel. Frac.	0.22	
	(0.44)	
Oil exporter	−0.10	
	(0.22)	
Log(Mountainous terrain)	0.06	
	(0.06)	
Log(Population)	0.16	
	(0.09)	
Country fixed effects	no	yes
Country-specific time trend	yes	yes
R2	0.401	0.535
RMSE	0.34	0.30
Num.Obs.	743	743

(j) Finally, suppose we were really interested in the effects of democracy and ethnic fractionalization on conflict. Choose a set of controls and fit your model with the following estimators

- A fixed effects estimator that allows for time-invariant effects

Table 6: First-stage F statistics

	Model 5	Model 6
Growth	5.74	4.76
Lagged growth	3.93	3.86

Table 7: GMM results

	Est	St. Err.
Growth	0.258	1.225
Democracy	-0.005	0.005
Population	0.173	0.983
Sargan Test	5.4 ($p \approx 0.02$)	

(e.g., a Hausman-Taylor-like estimator, with or without random effects)

- Munlak (with or without the random effects)

Interpret your results. What if anything do you find?

Table 8: Working with time invariant covariates

	HT	Mundlak
Democracy	-0.01	-0.01
	(0.01)	(0.01)
Eth. Frac.	-0.02	-0.02
	(1.51)	(0.19)
Lagged growth	-0.17	-0.17
	(0.21)	(0.21)
Log(Population)	0.08	0.08
	(0.99)	(1.01)
R2	0.100	0.155
Num.Obs.	742	742

The results of the two estimators (without random effects) are presented in Table 8. As expected, they are fairly similar. A 1 point increase in democracy is associated with a 1 percentage point decrease in the probability of civil conflict, on average. The only thing that standard out here is that the standard error on ethnic fractionalization is much higher in the Hausman Taylor estimator.