Advanced Applied Econometrics

Christophe Nozaradan

May 2021, KULeuven

1 Introduction

In this report, we perform a regression analysis with panel data. The economic question is related to economic growth theory, where one seeks to measure the relationship between GDP growth and relevant determinants. For the purpose of applying selected estimation methods and perform related specification tests, we first restrict this study to static panel models, even if dynamic models should be preferred[1]. In the last section, a dynamic panel model is also considered.

2 Data

We use the same data as in article [1]. We refer to Appendix A from this article for a complete description. We briefly describe the variables of interest here:

- `diffy`, growth of real GDP per capita (2000 US dollars at PPP). Source: PWT 6.2.
- `lny0`, logarithm of initial real GDP per capita (2000 US dollars at PPP). Source: PWT 6.2.
- `lni2`, logarithm of real investment as ratio to GDP (2000 US dollars at PPP). Source: PWT 6.2.
- `lnpopgr`, logarithm of annual population growth rate plus 0.05. Source: PWT 6.2.
- `lfexp2`, life expectancy at birth (total) with filled in years. Source: World Development Indicators.
- `lninfl`, logarithm of one plus the inflation rate. Source: International Financial Statistics (IMF).
- `d_debtgdp`, debt categorical variable: classification based on percentiles of external debt as percentage of GDP (0 if unreported, 1 if below 25th, 2 if between 25th and 49th, 3 if between 50th and 74th, 4 if above75th). Source: World Development Indicators (World Bank) and calculations from authors of [1].
- `lnindexpwt`, logarithm of index of exchange rate over/undervaluation. Source: PWT 6.2. and calculations from the authors of [1].
- `open2`, exports plus Imports as share of GDP (2000 US dollars at PPP). Source: PWT 6.2.

The panel data contains 1602 records, for 178 countries and 9 time periods (from 1965 to 2005 with a 5 year gap). There are missing values for the variable `diffy` (423 records out of 1602) which is selected as dependent variable. The panel is therefore unbalanced. We select the remaining variables as regressors.

[1] If the underlying data generating process is dynamic, then ignoring the lag of the dependent variable as regressor will create serial correlation in the error terms and other problems that we will try to address.

3 Static panel models with unobserved heterogeneity across entities

3.1 Global assumptions

The population model is postulated as follows:

$$\begin{aligned}\text{diffy}_{it} = \beta_0 + \beta_1 \cdot \text{lny0}_{it} + \beta_2 \cdot \text{lni2}_{it} + \beta_3 \cdot \text{lnpopgr}_{it} + \beta_4 \cdot \text{lfexp2}_{it} + \beta_5 \cdot \text{lninfl}_{it} \\ + \beta_6 \cdot \text{d\_debtgdp}_{it} + \beta_7 \cdot \text{lnindexpwt}_{it} + \beta_8 \cdot \text{open2}_{it} + a_i + \epsilon_{it} \ ,\end{aligned}$$

where β_0,..., β_8 are unknown constants, a_i are the unobserved heterogeneity and ϵ_{it} are the error terms. This model can be seen as a transformation of the equation $y_{it} = \alpha y_{it-1} + x'_{it}\beta + a_i + \epsilon_{it}$, where y_{it} denotes the logarithm of real GDP per capita and x_{it} the vector of regressors for entity i and time period t. One then subtracts y_{it-1} on both sides. On the LHS, we obtain our dependent variable $\Delta y_{it} := y_{it} - y_{it-1}$ (`diffy`$_{it}$). On the RHS, our restriction to static models in this section makes us drop the lagged term Δy_{it-1}. One further assumes that the regressors are not perfectly multicollinear, that the variables are iid across entities and that strict exogeneity holds: $E[e_{it}|x_{i1}, ..., x_{iT}] = 0$, where x_{it} represents the vector of regressors.

3.2 Fixed effects with homoskedastic errors (1)

The fixed effects population model can be written as follows:

$$\begin{aligned}\text{diffy}^*_{it} = \beta_1 \cdot \text{lny0}^*_{it} + \beta_2 \cdot \text{lni2}^*_{it} + \beta_3 \cdot \text{lnpopgr}^*_{it} + \beta_4 \cdot \text{lfexp2}^*_{it} + \beta_5 \cdot \text{lninfl}^*_{it} \\ + \beta_6 \cdot \text{d\_debtgdp}^*_{it} + \beta_7 \cdot \text{lnindexpwt}^*_{it} + \beta_8 \cdot \text{open2}^*_{it} + \epsilon^*_{it} \ ,\end{aligned}$$

where $*$ denotes that a variable is demeaned, e.g. $\text{diffy}^*_{it} = \text{diffy}_{it} - \overline{\text{diffy}}_i$. The assumption of strict exogeneity should hold, but the unobserved entity effect a_i may be correlated with the regressors.

The above model is estimated with OLS. The results are available in column (1) from table 1. All coefficients are significant at the one percent level. The null hypothesis of no unobserved heterogeneity is rejected since we obtained F(150, 706) = 2.54. This implies that the pooled panel model estimated with pooled OLS should be discarded.

3.3 First difference with homoskedastic errors (2)

The first difference population model can be written as follows:

$$\begin{aligned}\Delta\text{diffy}_{it} = \beta_1 \cdot \Delta\text{lny0}_{it} + \beta_2 \cdot \Delta\text{lni2}_{it} + \beta_3 \cdot \Delta\text{lnpopgr}_{it} + \beta_4 \cdot \Delta\text{lfexp2}_{it} + \beta_5 \cdot \Delta\text{lninfl}_{it} \\ + \beta_6 \cdot \Delta\text{d\_debtgdp}_{it} + \beta_7 \cdot \Delta\text{lnindexpwt}_{it} + \beta_8 \cdot \Delta\text{open2}_{it} + \Delta\epsilon_{it} \ ,\end{aligned}$$

where Δ denotes first time difference operator, e.g. $\Delta\text{diffy}_{it} = \text{diffy}_{it} - \text{diffy}_{it-1}$. The assumption of strict exogeneity should hold, but in differences.

The above model is estimated with OLS. If all required assumptions hold, the results (1) and (2) should be similar. If not, it is likely that the assumption of strict exogeneity may be violated. From the output (see table 1), the results (1) and (2) differ in magnitude, especially for `lny0`, `lfexp2` and `open2`. However, the sign of the coefficients remain consistent. We can further test the assumption of strict exogeneity following Wooldridge's approach.

3.4 Testing for strict exogeneity (Wooldridge's approach)

We apply Wooldridge's approach to test the assumption of strict exogeneity. For fixed effects, this amounts to include the lead of the regressors in the regression. We cannot use the lead of variable `lny0` since, by definition, we have $\text{lny0}_{t+1} - \text{lny0}_t = \text{diffy}_t$. Using `F2.lny0` instead and adding the lead of regressors of `lni2`, `lnpopgr` and `lninfl`, we obtain:

```
------------------------------------------------------------------------------
       diffy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lny0 |   -.525331   .0275092   -19.10   0.000    -.5793809   -.4712811
        lni2 |    .012616   .0224093     0.56   0.574    -.0314138    .0566458
     lnpopgr |  -.1372094   .0379273    -3.62   0.000    -.2117289     -.06269
      lfexp2 |   .0042033   .0021765     1.93   0.054    -.0000731    .0084796
      lninfl |  -.0769027   .0208538    -3.69   0.000    -.1178763   -.0359291
   d_debtgdp |  -.0195392   .0093094    -2.10   0.036    -.0378304   -.0012481
  lnindexpwt |    .061836   .0234932     2.63   0.009     .0156767    .1079953
       open2 |   .1222439   .0293858     4.16   0.000     .0645068     .179981
             |
        lny0 |
         F2. |   .3941541   .0330789    11.92   0.000     .3291609    .4591474
             |
        lni2 |
         F1. |  -.0453716   .0253058    -1.79   0.074    -.0950924    .0043492
             |
     lnpopgr |
         F1. |   .1316884    .036341     3.62   0.000     .0602857    .2030912
             |
      lninfl |
         F1. |     .04567   .0203666     2.24   0.025     .0056538    .0856862
             |
       _cons |   .9043885   .2430112     3.72   0.000     .4269207    1.381856
------------------------------------------------------------------------------
```

The results show significant coefficients at 1% level for `F2.lny0` and `F.lnpopgr` and at 5% level for `F.lninfl`. This indicates that the strict exogeneity assumption may not hold.

3.5 FE-IV regression with homoskedastic errors (3)

In an attempt to solve the endogeneity problem for `lny0`, `lnpopgr` and `lninfl`, we run a panel IV regression with lag variables as instruments (including `L4.lny0`). The results obtained for the first stage regression show that the instruments are relevant (including for `L4.lny0`). Since they are lagged variables, they are valid instruments unless the error terms are serially correlated (which is a possibility here because we restrict ourselves to static panel models). The results are available in column (3) of table 1.

In an attempt to obtain more efficient estimators, we now try a random effects model.

3.6 RE-IV regression with homoskedastic errors (4)

The random effects population model can be written as follows:

$$\begin{aligned}\text{diffy}_{it} = \beta_0 + \beta_1 \cdot \text{lny0}_{it} + \beta_2 \cdot \text{lni2}_{it} + \beta_3 \cdot \text{lnpopgr}_{it} + \beta_4 \cdot \text{lfexp2}_{it} + \beta_5 \cdot \text{lninfl}_{it} \\ + \beta_6 \cdot \text{d\_debtgdp}_{it} + \beta_7 \cdot \text{lnindexpwt}_{it} + \beta_8 \cdot \text{open2}_{it} + \nu_{it} \ ,\end{aligned}$$

where $\nu_{it} = a_i + \epsilon_{it}$. In contrast with the fixed effects, here the strict exogeneity assumption implies that $\text{Cov}(a_i, x_{it}) = 0$, where x_{it} represents the vector of regressors.

The above model is estimated with GLS. We apply the same IV setting as in the fixed-effect case. The results are presented in column (4) in table 1.

We now perform a Hausman test to decide which of the FE or RE models should be preferred.

3.7 Hausman test (for assumed homoskedastic errors)

We use the Stata command `hausman` (with the `sigmamore` option) to obtain:

```
                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |    FEIVhom      REIVhom       Difference          S.E.
-------------+------------------------------------------------------------
         lny0 |   -.1223698    -.0759986       -.0463712        .0542683
      lnpopgr |   -.2622947     .0831795       -.3454742               .
       lninfl |   -.3354931    -.0751388       -.2603543        .0873603
         lni2 |    .1122226     .0923116        .0199111        .0354679
      lfexp2 |    -.003215     .0037024       -.0069175        .0024672
   d_debtgdp |   -.0402303    -.0409642        .0007338         .010269
  lnindexpwt |    .0892944     .0777736        .0115208        .0249031
       open2 |    .1047798     .0382873        .0664925        .0342716
------------------------------------------------------------------------------
                        b = consistent under Ho and Ha; obtained from xtivreg
         B = inconsistent under Ha, efficient under Ho; obtained from xtivreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       49.94
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

The Hausman test rejects the null hypothesis. Therefore the results obtained with the fixed effect estimator should be preferred (since based on weaker assumptions).

3.8 FE-IV regression with heteroskedastic errors/HAC correction (5)

The assumptions of a fixed effects model leave open the possibility that the regressors and/or the error terms are serially correlated. If the error terms are serially correlated, we know that the usual heteroskedasticity-robust standard errors are wrong. We can correct these with so-called "HAC" standard errors. One type of such HAC standard errors are "clustered standard errors". The corresponding results are shown in column (5) of table 1. Comparing the results (5) and (3), we observe slight changes in the standard errors. The estimates for the coefficients of `lny0` and `lnpopgr` are more precise in (5). The remaining ones, however, loose significance.

4 Dynamic panel models with unobserved heterogeneity across entities/time

The postulated model is now written as follows:

$$\begin{aligned}\text{diffy}_{it} = \beta_0 + \gamma \cdot \text{diffy}_{it-1} + \beta_1 \cdot \text{lny0}_{it} + \beta_2 \cdot \text{lni2}_{it} + \beta_3 \cdot \text{lnpopgr}_{it} + \beta_4 \cdot \text{lfexp2}_{it} + \beta_5 \cdot \text{lninfl}_{it} \\ + \beta_6 \cdot \text{d\_debtgdp}_{it} + \beta_7 \cdot \text{lnindexpwt}_{it} + \beta_8 \cdot \text{open2}_{it} + a_i + b_t + \epsilon_{it} \ ,\end{aligned}$$

Note that we also include time effects b_t in the model.

We use a two-step system GMM estimation with the Stata command `xtabond2`, with `lny0`, `lnpopgr` and `lninfl` as endogenous variables. We use as instruments the corresponding lag variables up to lag 3. As strictly exogenous variables, we use the year dummies, `lni2`, `lfexp2`, `ltoted2`[2] and `d_debtgdp` (although their exogenous character can be debated). The estimated coefficients and their significance level are shown in column (6) in table 1. We see that the lag of the dependent variable included as regressor is not significantly different than zero. The coefficients of `lny0` and `lni2` are significant at the 1% level.

The other relevant results are shown below:

```
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -4.71  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -1.86  Pr > z =  0.063
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(86)   = 139.02  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(86)   =  87.88  Prob > chi2 =  0.423
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  GMM instruments for levels
    Hansen test excluding group:     chi2(57)   =  69.40  Prob > chi2 =  0.126
    Difference (null H = exogenous): chi2(29)   =  18.49  Prob > chi2 =  0.934
  gmm(L.diffy, lag(1 .))
    Hansen test excluding group:     chi2(55)   =  60.47  Prob > chi2 =  0.285
    Difference (null H = exogenous): chi2(31)   =  27.41  Prob > chi2 =  0.651
  gmm(lny0 lnpopgr lninfl, lag(2 3))
    Hansen test excluding group:     chi2(26)   =  24.76  Prob > chi2 =  0.532
    Difference (null H = exogenous): chi2(60)   =  63.12  Prob > chi2 =  0.367
  iv(lni2 lfexp2 ltoted2 d_debtgdp yr1 yr2 yr3 yr4 yr5 yr6 yr7 yr8)
    Hansen test excluding group:     chi2(75)   =  77.08  Prob > chi2 =  0.412
    Difference (null H = exogenous): chi2(11)   =  10.80  Prob > chi2 =  0.460
```

Using [2] as guidance to interpret this output, we see that the Hansen test of overidentifying restrictions is clearly rejected (p-value=0.423) in favor of the null hypothesis that the instruments are valid (exogenous). The test on lack of second order autocorrelation does not seem to be conclusive, although it does not reject the null hypothesis of no second order autocorrelation in the error terms at the 5% rejection level. The difference-in-Hansen tests of exogeneity of instruments subsets all favor the null hypothesis of instrument validity.

Finally, we tested the null hypothesis that the year dummies included in the model are all equal to zero. It was clearly rejected, indicating that they add significance to the model.

[2] This is the logarithm of the total average stock of years of primary and secondary education. Source: see reference in [1].

Table 1: Regression table

	(1) diffy	(2) D.diffy	(3) diffy	(4) diffy	(5) diffy	(6) diffy
lny0	-0.273***		-0.122	-0.0760***	-0.122*	-0.109***
	(0.0193)		(0.0773)	(0.0168)	(0.0527)	(0.0323)
lni2	0.0726***		0.112*	0.0923***	0.112	0.0824***
	(0.0210)		(0.0534)	(0.0167)	(0.0668)	(0.0161)
lnpopgr	-0.170***		-0.262**	0.0832	-0.262***	-0.157
	(0.0379)		(0.0927)	(0.181)	(0.0611)	(0.0908)
lfexp2	0.00485**		-0.00322	0.00370*	-0.00322	0.00505*
	(0.00168)		(0.00403)	(0.00163)	(0.00461)	(0.00234)
lninfl	-0.130***		-0.335*	-0.0751	-0.335	-0.0648
	(0.0215)		(0.140)	(0.0549)	(0.191)	(0.0396)
d_debtgdp	-0.0331***		-0.0402*	-0.0410***	-0.0402	-0.0425***
	(0.00843)		(0.0193)	(0.00983)	(0.0233)	(0.0114)
lnindexpwt	0.0763**		0.0893*	0.0778***	0.0893	0.109***
	(0.0237)		(0.0437)	(0.0203)	(0.0724)	(0.0273)
open2	0.181***		0.105*	0.0383*	0.105	0.0583
	(0.0242)		(0.0530)	(0.0184)	(0.0872)	(0.0304)
D.lny0		-0.666***				
		(0.0326)				
D.lni2		0.0446				
		(0.0256)				
D.lnpopgr		-0.146***				
		(0.0335)				
D.lfexp2		0.0131***				
		(0.00292)				
D.lninfl		-0.0942***				
		(0.0200)				
D.d_debtgdp		-0.0315**				
		(0.0118)				
D.lnindexpwt		0.0811**				
		(0.0290)				
D.open2		0.280***				
		(0.0384)				
L.diffy						0.0247
						(0.0484)
yr2						0.0620
						(0.0327)
yr3						0.0171
						(0.0290)
yr4						0.0423
						(0.0224)
yr5						-0.0341*
						(0.0167)
yr6						0.0462*
						(0.0194)
yr7						0.0180
						(0.0155)
yr8						0.0433***
						(0.0128)
_cons	2.160***		1.586**	0.144	1.586***	0.780**
	(0.180)		(0.593)	(0.450)	(0.445)	(0.253)
N	865	713	479	479	479	693
adj. R^2	0.219	0.472				

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

References

[1] Alin Mirestean and Charalambos Tsangarides. Growth determinants revisited using limited-information bayesian model averaging. *Journal of Applied Econometrics*, 31, 06 2015.

[2] David Roodman. How to do xtabond2: An introduction to difference and system GMM in Stata. *Stata Journal*, 9(1):86–136, March 2009.