Advanced Applied Econometrics - Paper

Vincent Buekers - r0754046

1 Introduction

This paper examines to what extent infrastructures contributes to the Gross Domestic Product (GDP) of a given country. To that end, components such as physical capital, labour, roads and phone lines are taken into account. Policy decisions regarding investments in public infrastructure depend on how they are estimated to contribute to aggregate output and productivity. As such, a statistically sound model should allow policy makers to more effectively decide on their infrastructure investments such that they suit the national economic interests.

The panel data sample used in the analysis consists of 3,976 observations across 97 countries, where each country has been observed from 1960 up to and including 2000. Both static and dynamic models will be examined in order to gain insights from these panel data. Note that STATA is used to conduct these analyses, of which the code has been added to appendix.

For overview, the following 15 variables are included in the sample:

- **year**: Year
- **cid**: Country identifier (Number)
- **code**: Country identifier (Text)
- **country**: Country name
- **y**: Gross Domestic Product (GDP)
- **k**: Physical capital
- **kbc30**: Physical capital (computed via 'backcasting')
- **lwdi**: Labour
- **secondary**: Years of secondary education
- **egc**: electricity generating capacity
- **mlines**: Main phone lines
- **troads**: total roads
- **cells**: cell phones
- **proads**: Paved roads
- **rails**: Railroads

Note that in the subsequent analyses, the physical capital as estimated by Calderon & Moral-Benito (2015), kbc30, will be used instead of the observed physical capital (k) due to a large amount of missing values.

2 Static Models

2.1 Pooled OLS

This first part of the paper consists of static models for analyzing panel data. Firstly, a simple Pooled OLS is estimated. This model can be summarized as (1) and it only holds under the assumption of contemporaneous exogeneity of X. The flexibility of this assumption regarding the relationships over time does allow for the inclusion of lagged values of y (e.g. $y_{i,t-1}$)

$$Y_{it} = X_{it}^T\beta + \varepsilon_{it} \tag{1}$$

Where $E(\varepsilon_{it}|x_{it}) = 0$

Given the non-restrictive nature of this Pooled model, it serves mainly to gain some preliminary insights and to identify whether or not there are problems with heteroskedasticity. As it turns out (Figure 1), the errors terms are not homoskedastic since the assumption of constant variance is rejected ($p < 0.0001$).

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y

         chi2(1)      = 39634.47
         Prob > chi2  =   0.0000
```

Figure 1: Breusch-Pagan Test

Consequently, it is necessary to use a robust covariance matrix. This results in the following output (Figure 2). The covariates *kbc30, lwdi, mlines and proads* are considered significant w.r.t. the country GDP, as well as the lagged value of the GDP itself. For now, it seems as though the main phone lines and paved roads are important infrastructure resources in terms of its contribution the yearly national income.

y	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
y						
L1.	1.057691	.0134618	78.57	0.000	1.031298	1.084084
kbc30	-.0132612	.0038298	-3.46	0.001	-.0207699	-.0057525
lwdi	125.7214	20.13759	6.24	0.000	86.24006	165.2027
secondary	-3.94e+08	3.94e+08	-1.00	0.318	-1.17e+09	3.79e+08
egc	-36577.25	106440.6	-0.34	0.731	-245262.2	172107.7
mlines	1113.532	410.1708	2.71	0.007	309.3602	1917.703
troads	-812.5956	4840.592	-0.17	0.867	-10302.95	8677.76
cells	202.0038	328.4665	0.61	0.539	-441.9801	845.9877
proads	-33602.49	9949.036	-3.38	0.001	-53108.34	-14096.63
rails	82502.82	76167.86	1.08	0.279	-66830.17	231835.8
_cons	-6.29e+08	4.16e+08	-1.51	0.130	-1.44e+09	1.86e+08

Figure 2: Pooled OLS

2.2 Unobserved Heterogeneity: Fixed Effects

Secondly, a fixed effects regression is conducted in order to assess the within group variation. As such, the variation between countries is ignored for now. To that end, the data is transformed by subtracting the mean over time per group. This results in the following functional form (2).

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)^T \beta + u_{it} \tag{2}$$

Where Y is the country GDP in a certain year $\in [1960, 2000]$ and X the design matrix containing an intercept and the variables measuring the physical capital, human capital and infrastructure resources. Under this fixed effects model, it is important that the errors u_{it} are uncorrelated with the dependent variables for all time periods, known as the strict exogeneity assumption. Formally, this is denoted as $E(u_{it}|X_{it}, c_i)$. As such, the inclusion of lagged values of the GDP is no longer possible since this would violate the assumption by construction. The results obtained for (2) are displayed below (Figure 3).

y	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
kbc30	.1732702	.0228651	7.58	0.000	.1278832	.2186571
lwdi	4362.078	462.0561	9.44	0.000	3444.903	5279.252
secondary	-6.32e+09	4.85e+09	-1.30	0.196	-1.59e+10	3.30e+09
egc	-287968.4	383100.8	-0.75	0.454	-1048417	472480.7
mlines	13194.42	2961.248	4.46	0.000	7316.383	19072.45
troads	128473.2	37095.43	3.46	0.001	54839.41	202107.1
cells	862.6945	3448.368	0.25	0.803	-5982.262	7707.651
proads	57200.95	120300.4	0.48	0.636	-181593.4	295995.3
rails	-7772356	1553047	-5.00	0.000	-1.09e+07	-4689581
_cons	6.63e+10	1.62e+10	4.10	0.000	3.42e+10	9.84e+10
sigma_u	3.443e+11					
sigma_e	3.716e+10					
rho	.98848787	(fraction of variance due to u_i)				

Figure 3: Fixed effects regression in levels, robust standard errors

As indicated by the test for unobserved heterogeneity (F-test for all $H_0 : u_i = 0, \forall i$), the null hypothesis is rejected ($p < 0.0001$). This means that, apart from the included variables, there is still some country-specific effect not accounted for by the model. The covariates considered significant w.r.t. the country GDP are *kbc30, lwdi, mlines, troads and rails.* Whereas paved roads where considered important contributions in the OLS model, now the total amount of roads and rails are considered as such.

It is recommended to conduct a first-differences model to assess the similarity with the fixed effects model in levels. The unobserved heterogeneity model in first differences is also a model on fixed effects. As such, the strict exogeneity assumption should still hold, albeit in differences. The functional form is as follows.

$$\Delta Y_{it} = \Delta X_{it}^T \beta + u_{it} \tag{3}$$

Where $E(\Delta u_{it}|\Delta X_{it}) = 0$

The results corresponding to (3) are displayed below (Figure 4). It is immediately clear that the results are quite different from the fixed effects model in levels. The total amount of roads and rails are no longer considered significant, while cell phone and paved roads are now significant.

D.y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
kbc30						
D1.	.3417079	.0059348	57.58	0.000	.3300723	.3533434
lwdi						
D1.	2169.757	182.9876	11.86	0.000	1810.996	2528.519
secondary						
D1.	-1.04e+10	5.47e+09	-1.89	0.058	-2.11e+10	3.66e+08
egc						
D1.	-350129.3	133044.9	-2.63	0.009	-610974.1	-89284.49
mlines						
D1.	5493.315	371.2024	14.80	0.000	4765.544	6221.086
troads						
D1.	18248.65	11950.27	1.53	0.127	-5180.768	41678.07
cells						
D1.	-1384.904	249.8798	-5.54	0.000	-1874.812	-894.9949
proads						
D1.	294681.4	26297.73	11.21	0.000	243122.7	346240.2
rails						
D1.	219916.7	228903.1	0.96	0.337	-228865.4	668698.8

Figure 4: Fixed effects, first differences

This difference between the model in levels and in differences is troublesome since this points to a violation of the strict exogeneity assumption. Consequently, the assumptions discussed above for both the model in levels and in differences need to be appropriately tested for.

2.3 Testing Strict Exogeneity

As a next step, the strict exogeneity assumption is tested for the fixed effects model both in levels and in differences. In practice, this does require two different procedures. Testing the assumption for the model in levels involves a fixed effects regression where the first-order leads of the regressors are included. On the other hand, the model in differences is adapted by including the levels of the variables, in addition to the first differences. As it turns out, both tests reject the null-hypothesis for strict exogeneity ($p < 0.0001$). These F-test results on global significance imply that the unobserved heterogeneity structure exists over time. Consequently, the errors of the model are not uncorrelated with the variables over time, rendering the assumption violated. As such, the concerns postulated above were not unjustified. Note that the exact results are provided in the supplemental log-files of this paper.

2.4 Unobserved Heterogeneity: Random Effects

Finally, it could be interesting te conduct a model on random effects in order to treat the unobserved heterogeneity. Now the between group variation is no longer ignored, as it does no longer subtract the group means in the estimation. This can be easily seen in its specification (4).

$$Y_{it} = X_{it}^T \beta + c_i + \varepsilon_{it} \tag{4}$$

Where $Cov(X_{it}, c_i) = 0$

The results corresponding to (4) are displayed below (Figure 5). Once again, robust standard errors are used due to the recurrent problem of heteroskedasticity. the covariates *kbc30, lwdi, mlines, troads and rails* are considered significant, as they were in the fixed effects model in levels. Additionally, also the years of secondary education (i.e. human capital) is now of apparent importance.

y	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
kbc30	.1697082	.0496369	3.42	0.001	.0724216	.2669949
lwdi	2699.023	327.3812	8.24	0.000	2057.368	3340.678
secondary	-1.23e+10	5.81e+09	-2.11	0.035	-2.37e+10	-8.92e+08
egc	753337.4	531820.5	1.42	0.157	-289011.6	1795686
mlines	13632.04	2988.317	4.56	0.000	7775.048	19489.03
troads	283997.5	94722.07	3.00	0.003	98345.69	469649.4
cells	1401.382	3087.403	0.45	0.650	-4649.816	7452.58
proads	67048.01	214003.7	0.31	0.754	-352391.6	486487.6
rails	-4272942	1192514	-3.58	0.000	-6610227	-1935657
_cons	2.06e+10	1.46e+10	1.42	0.157	-7.92e+09	4.91e+10
sigma_u	4.576e+10					
sigma_e	3.716e+10					
rho	.60264988	(fraction of variance due to u_i)				

Figure 5: Random effects, robust standard errors

Lastly, these random effects can also be tested for by a test from Breusch and Pagan. As indicated by this test, $H_0 : Var(c) = 0$ is rejected, meaning there are indeed random effects present between the countries. Since robust standard errors are used, the Hausman test can not be conducted.

```
Breusch and Pagan Lagrangian multiplier test for random effects

        y[cid,t] = Xb + u[cid] + e[cid,t]

        Estimated results:
                        |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                       y |   4.45e+23       6.67e+11
                       e |   1.38e+21       3.72e+10
                       u |   2.09e+21       4.58e+10

        Test:   Var(u) = 0
                             chibar2(01) = 16636.61
                          Prob > chibar2 =   0.0000
```

Figure 6: Breusch-Pagan Lagrangian Multiplier test for Random Effects

3 Dynamic Models

In this section, lagged values of the dependent variable Y are to be included in the model, allowing for the estimation of parameters that reflect a dynamic relationship between the GDP and the (human) capital and infrastructure variables. For a single observation, this corresponds to equation (5) below.

$$y_{it} = \gamma y_{i,t-1} + x_{it}\beta + \alpha_i + \varepsilon_{it} \tag{5}$$

Where $E(y_{i,t-1}\varepsilon_{it}) = 0$

Applying the estimators used above in the static models would result in inconsistent estimates due to violation of the assumption that the lagged dependent variable should be uncorrelated with the error term. Dynamic methods such as Instrumental Variable (IV) estimation or Generalized method of moments (GMM) estimation offer a solution to this problem.

3.1 IV Estimation

The Instrumental Variable (IV) procedure starts by taking the first differences of equation (5), through which it eliminates the country-specific effect α. This does not yet solve the assumption violation however. As suggested by Anderson and Hsiao (1981), taking $\Delta y_{i,t-2}$ or $y_{i,t-2}$ as instruments is a way to circumvent this issue, given that the error terms ε are not autocorrelated. Since T is large enough in this sample, the lagged difference of y is used.

The results obtained for the IV regression using $\Delta y_{i,t-2}$ are given in Figure 8 below. Note that in order to protect against issues of heteroskedasticity, robust standard errors have again been used in the estimation. Indeed, the non-robust estimation suspiciously reported all variables to be highly significant. At the significance level of $\alpha = 0.05$, the variables kbc30, lwdi and mlines are considered significant predictors w.r.t. the country GDP. As it turns out, the dynamic relationship involving the addition of $y_{i,t-1}$ is not considered significant ($p > 0.05$).

D.y	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
y						
LD.	-.879123	.8037454	-1.09	0.274	-2.454435	.6961889
kbc30						
D1.	.675879	.3142945	2.15	0.032	.0598731	1.291885
lwdi						
D1.	3895.08	1738.944	2.24	0.025	486.8124	7303.347
secondary						
D1.	-2.23e+10	1.38e+10	-1.61	0.106	-4.93e+10	4.77e+09
egc						
D1.	-626241.6	938664.7	-0.67	0.505	-2465991	1213507
mlines						
D1.	7879.711	3275.181	2.41	0.016	1460.474	14298.95
troads						
D1.	101010.2	72160.6	1.40	0.162	-40421.96	242442.4
cells						
D1.	-3363.725	2359.838	-1.43	0.154	-7988.922	1261.472
proads						
D1.	316932.2	196317.5	1.61	0.106	-67843.07	701707.5
rails						
D1.	1194138	2095691	0.57	0.569	-2913342	5301617
_cons	5.93e+07	2.76e+08	0.21	0.830	-4.81e+08	6.00e+08

Figure 7: IV estimation using $\Delta y_{i,t-2}$

Although the IV estimation is consistent, it lacks efficiency as it does not use all available information of the model. To that end, one can also rely on GMM estimation.

3.2 GMM Estimation

The final procedure conducted on the infrastructure data involves Generalized-Method-of-Moments estimation, which in the context of panel data is referred to as the Arellano-Bond estimator (Arellano and Bond, 1991). There are however, different alternatives of the estimator depending on the assumptions regarding the correlation between x_{it} and ε_{is} $(t \neq s)$. The case of strictly exogenous variables seems somewhat unrealistic in the setting of infrastructure investments and GDP. This would mean that the covariates x_{it} are not allowed to correlate with ε_{is} whatsoever. Alternatively, it would make more sense to allow for x_{it} to correlate with past errors (i.e. $s < t$), yet not with contemporary or future error terms $(s > t)$. x_{it} are described as predetermined in this case.

An additional consideration is the choice between one-step GMM and two-step GMM estimation, which are both consistent but vary in terms efficiency in presence of heteroskedasticity. Since two-step GMM is favorable when ε_{it} are heteroskedastic, it should be the better choice given the previously encountered issues on constant variance. In line with the other estimations, a finite sample correction has also been applied in order to obtain robust standard errors. Lastly, it should be noted that a restriction has been imposed on the amount of lags used as instruments. The maximum lag length has been set to three in order to keep the number of instruments reasonable relative to the amount of observations, even more so because of the high dimensionality of the model (i.e. a lot of explanatory variables).

The output obtained for the model outlined above is given in Figure 9.

```
Dynamic panel-data estimation, two-step difference GMM

Group variable: cid                          Number of obs      =      3783
Time variable : year                         Number of groups   =        97
Number of instruments = 1765                 Obs per group: min =        39
Wald chi2(11) =  1.02e+08                                   avg =     39.00
Prob > chi2   =     0.000                                   max =        39
```

y	Coef.	Corrected Std. Err.	z	P>|z|	[95% Conf.	Interval]
y						
L1.	1.026274	.0567011	18.10	0.000	.9151417	1.137406
kbc30	-.0134364	.0104283	-1.29	0.198	-.0338756	.0070027
lwdi	844.9458	184.9942	4.57	0.000	482.3638	1207.528
secondary	2.84e+09	3.82e+09	0.74	0.458	-4.65e+09	1.03e+10
egc	-196645.1	94813.17	-2.07	0.038	-382475.5	-10814.7
mlines	1413.81	645.5957	2.19	0.029	148.4653	2679.154
troads	-40145.74	14660.9	-2.74	0.006	-68880.57	-11410.91
cells	523.1697	597.6609	0.88	0.381	-648.2243	1694.564
proads	-7388.233	32610.64	-0.23	0.821	-71303.91	56527.44
rails	-577608.9	380346.4	-1.52	0.129	-1323074	167856.4
year	-2.07e+08	8.82e+07	-2.35	0.019	-3.80e+08	-3.45e+07

Figure 9: Two-step GMM with predetermined variables and finite sample correction

Clearly, the GMM results are not quite in agreement with those of the IV estimation (Figure 7). Whereas earlier, the lagged value of the GDP was not significant, it is now. Among the remaining explanatory variables *kbc30* is no longer significant, while *egc and troads* are. The dynamic models only seem to agree on *lwdi* and *mlines*.

Since the validity of these instruments is crucial for the consistency of the GMM estimator, this also needs to be appropriately tested for. The two alternatives for such a test on overidentifying restrictions include the Sargan and the Hansen test, the latter of which is appropriate for the two-step GMM. Moreover, it is also consistent in presence of heteroskedasticity. The results for this test statistic are $\chi^2 = 95.21, p = 1$. As such, H_0 : instruments are valid, is not rejected. Assuming the test is not severely affected by the amount of instruments, the model specification can be deemed appropriate by the validity of instruments. Secondly, the consistency of the GMM estimator also depends on ε_{it} not being autocorrelated. Since we are working in differences, it means that first order autocorrelation is allowed, yet second order is not. This involves testing $H_0 : E(\Delta\varepsilon_{it}\Delta\varepsilon_{i,t-2}) = 0$. The results are the following:

```
Arellano-Bond test for AR(1) in first differences: z =  -1.90  Pr > z =  0.058
Arellano-Bond test for AR(2) in first differences: z =  -1.28  Pr > z =  0.202
```

Figure 10: Test on autocorrelation of ε_{it}

With regards to both the first-order and the second-order autocorrelation of ε_{it}, the null-hypothesis is not rejected at significance level $\alpha = 0.05$. Consequently, the consistency of the GMM estimation should now be assured as ε_{it} are not autocorrelated.

4 Conclusion

In summary, there are quite some differences across all the models. Most models point out the importance of physical capital (kbc30), as well as labour (lwdi). Human capital (secondary) and electricity generating capacity (egc) were not considered valuable contributions overall. Among the infrastructure investment variables, main phone lines (mlines) seemed to be most prominent. Furthermore, the total amount of roads (troads) and rails were also often considered valuable contributions.

5 Appendix

5.1 STATA Code

```
clear
capture log close
set more off

cd "/Users/Vincent/Desktop/AA Econometrics/Paper"
log using Paper_Vincent_Buekers.log, replace

* -----------------------------------------------------
* Do-file: Paper for Advanced Applied Econometrics 2019
* Vincent Buekers
*------------------------------------------------------

* Data
import delimited data-cmbs.txt

* Panel data specification
tsset cid year

* remove k due to lots of missing values.
* --> kbc30 will be used as physical capital variable
drop k

* info on panel data
xtdes

*-----------------------
* Descriptive statistics
*-----------------------

xtsum y kbc30 lwdi secondary egc mlines troads cells proads rails

*-------------------------
* Static Regression models
*-------------------------
*******
* OLS *
*******
reg y L.y kbc30 lwdi secondary egc mlines troads cells proads rails

* Test on homoskedasticity

    estat hettest

```

```
* robust OLS
        reg y L.y kbc30 lwdi secondary egc mlines troads cells proads rails, robust

*****************
* FIXED EFFECTS *
*****************

** within estimation
        xtreg y kbc30 lwdi secondary egc mlines troads cells proads rails, fe
        est store fixed

** additional goodness of fit criteria
        estat ic

** fitted values:

    ** linear prediction (x*beta_hat)
        predict hfe_xb, xb

    ** linear prediction including FE component (x*beta_hat + alphai_hat)
        predict hfe_xb_alphai, xbu

    ** prediction of FE (alphai_hat)
        predict hfe_alphai, u

    ** prediction of idiosyncratic error term (e_hat)
        predict hfe_e, e

    ** prediction of FE and idiosyncratic error term (alphai_hat + e_hat)
        predict hfe_alphai_e, ue


    list cid year y hfe_xb hfe_xb_alphai hfe_alphai hfe_e in 1/20


** within estimation with heteroskedasticity-robust standard errors:

    xtreg y kbc30 lwdi secondary egc mlines troads cells proads rails, fe robust


** First differences
        reg D.y D.kbc30 D.lwdi D.secondary D.egc D.mlines D.troads D.cells D.proads D.rails, nocons

*****************************
** test for strict exogeneity
*****************************
        * within
        xtreg y kbc30 F.kbc30 lwdi F.lwdi secondary F.secondary egc F.egc mlines F.mlines
```

```
        troads F.troads cells F.cells proads F.proads rails F.rails, fe

        * FD
        reg D.y D.kbc30 kbc30 D.lwdi lwdi D.secondary secondary D.egc egc D.mlines mlines
        D.troads troads D.cells cells D.proads proads D.rails rails, nocons

*****************
** Random effects
*****************

        xtreg y kbc30 lwdi secondary egc mlines troads cells proads rails, re
        est store random

** fitted values:

    ** linear prediction (x*beta_hat)
        predict hre_xb, xb

    ** linear prediction including FE component (x*beta_hat + alphai_hat)
        predict hre_xb_alphai, xbu

    ** prediction of FE (alphai_hat)
        predict hre_alphai, u

   ** prediction of idiosyncratic error term (e_hat)
        predict hre_e, e

    ** prediction of FE and idiosyncratic error term (alphai_hat + e_hat)
        predict hre_alphai_e, ue

list cid year y hre_xb hre_xb_alphai hre_alphai hre_e in 1/20


** Test on random effects

    xttest0

** RE estimation with heteroskedasticity-robust standard errors:

    xtreg y kbc30 lwdi secondary egc mlines troads cells proads rails, re robust
        hausman fixed .


*-----------------
* Dynamic models
*-----------------
********
** IV **
```

```
********

*Panel IV estimates (non robust SE's)
xtivreg y kbc30 lwdi secondary egc mlines troads cells proads rails (L.y = L2.y), fd

* Robust SE's using standard IV using DELTA(y_t-2) as instrument
ivregress 2sls D.y D.kbc30 D.lwdi D.secondary D.egc D.mlines D.troads D.cells D.proads
D.rails (L.D.y = L2.D.y)
ivregress 2sls D.y D.kbc30 D.lwdi D.secondary D.egc D.mlines D.troads D.cells D.proads
D.rails (L.D.y = L2.D.y), robust

*******
* GMM *
*******
* Two-step + finite sample correction + predetermined variables + laglimit = 3
xtabond2 y L.y kbc30 lwdi secondary egc mlines troads cells proads rails year, gmm(L.y)
gmm(kbc30, laglimits(1 3)) gmm(lwdi, laglimits(1 3)) gmm(secondary, laglimits(1 3)) gmm(egc, laglimits(1 3))
gmm(mlines, laglimits(1 3)) gmm(troads, laglimits(1 3)) gmm(cells, laglimits(1 3)) gmm(proads, laglimits(1 3))
gmm(rails, laglimits(1 3)) iv(year) noleveleq twostep robust

log close
```