Exercise Sheet 1: Estimation of Linear Panel Data Models

Review the Concepts and Proofs

1. Explain how panel data, at least in principle, solve the omitted variables problem that often plagues purely cross-sectional data. Which assumption do you have to make? Are they restrictive?

2. Discuss the strict exogeneity assumption. How can you check whether it holds?

3. Why is it typically a good idea to model time-invariant individual effects as being random as opposed to fixed? Think of a special case where it might be sensible to rather assume fixed effects.

4. Write down the variance matrix of the error components model.

5. Show that the POLS estimator is consistent. Find its asymptotic distribution.

6. Show that the FE estimator is consistent. Find its asymptotic distribution.

7. Write out the matrices $\mathbf{Q}_T$ and $\mathbf{J}_T$ for $T = 4$. Explain what they do if they are pre-multiplied to a $T \times K$ matrix of observations.

8. Show that $\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i'\mathbf{y}_i$.

9. Why is the FE variance estimator $\hat{\sigma}_u^2 = \frac{1}{N(T-1)-K}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^2$ normalized by $N(T-1)-K$ and not by $NT-K$? Does it matter asymptotically as $N \to \infty$?

10. Derive the matrix representation of the FE estimator.

11. Explain both the Wooldridge and the Swamy-Arora approach to estimate the variance components for the RE estimator. Which problem can arise? How is it typically dealt with?

12. Derive the matrix representation of the RE estimator.

13. Compare the POLS, FE, and RE estimators and their associated transformations. What are their advantages and disadvantages? How can you decide which estimator to use?

14. Explain the Hausman test. Find its asymptotic distribution.

Paper-pen exercises

1. Consider the matrices $\mathbf{J}_T = \boldsymbol{\iota}_T(\boldsymbol{\iota}_T'\boldsymbol{\iota}_T)^{-1}\boldsymbol{\iota}_T'$, $\mathbf{Q}_T = \mathbf{I}_T - \mathbf{J}_T$, $\mathbf{J} = \mathbf{I}_N \otimes \mathbf{J}_T$, and $\mathbf{Q} = \mathbf{I}_N \otimes \mathbf{Q}$, where $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones.

   (a) Show that $\mathbf{J}_T'\mathbf{J}_T = \mathbf{J}_T$, $\mathbf{Q}_T'\mathbf{Q}_T = \mathbf{Q}_T$, $\mathbf{J}_T\boldsymbol{\iota}_T = \boldsymbol{\iota}_T$, $\mathbf{Q}_T\boldsymbol{\iota}_T = \mathbf{0}$, and $\mathbf{Q}_T\mathbf{J}_T = \mathbf{0}$.

   (b) Show that the within-transformed regressors $\ddot{\mathbf{X}}_i$ and the between-transformed regressors $\bar{\mathbf{X}}_i$ are orthogonal to each other.

   (c) Show that $\mathbf{J}'\mathbf{J} = \mathbf{J}$, $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}$, and $\mathbf{Q}'\mathbf{J} = \mathbf{0}$.

2. Consider the error components model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_N \otimes \boldsymbol{\iota}_T)\mathbf{c} + \mathbf{u}$$

   with $\boldsymbol{\Omega} = \sigma_c^2\boldsymbol{\iota}_T\boldsymbol{\iota}_T' + \sigma_u^2\mathbf{I}_T$.

   (a) Show that the FE estimator is equivalent to the OLS estimator applied to the within-transformed equation.

   (b) Show that the between estimator is equivalent to the OLS estimator applied to the between-transformed equation.

   (c) Show that the RE estimator erroneously applied to the within-transformed equation,

$$\hat{\boldsymbol{\beta}}_{err} = \left(\ddot{\mathbf{X}}'[\mathbf{I}_N \otimes \hat{\boldsymbol{\Omega}}^{-1}]\ddot{\mathbf{X}}\right)^{-1}\ddot{\mathbf{X}}'[\mathbf{I}_N \otimes \hat{\boldsymbol{\Omega}}^{-1}]\ddot{\mathbf{y}},$$

   is identical to the FE estimator.

   (d) What is the result of the FE estimator applied to the between-transformed equation? Explain.

3. Acemoglu et al. (2008) analyze the effect of income on democracy.[1] They use a large country panel for 1960-2000 sampled at five-year intervals. Their baseline specification is

$$dem_{it} = \beta_1 dem_{i,t-1} + \beta_2 inc_{i,t-1} + \mu_t + c_i + u_{it}, \tag{1}$$

where dem_{it} denotes the democracy score of country i in period t (measured as the Freedom House Political Rights Index and scaled so that it is between zero and one, with one corresponding to the most democratic set of institutions), inc_{it} denotes log income per capita (in constant 1990 US dollars), and μ_t is a full set of year dummies.

   (a) A pooled OLS regression of dem_{it} on inc_{it} yields the results presented below. The variable "code_numeric" is a country identifier used when computing robust s.e.'s.

```
. reg dem inc if sample==1, cluster(code_numeric)

Linear regression                                   Number of obs =     960
                                                    F(  1,   151) =  414.02
                                                    Prob > F      =  0.0000
                                                    R-squared     =  0.4488
                                                    Root MSE      =  .26857

                                  (Std. Err. adjusted for 152 clusters in code_numeric)
------------------------------------------------------------------------------
             |               Robust
         dem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .2310104   .0113533    20.35   0.000     .2085785    .2534422
       _cons |  -1.339442   .0975336   -13.73   0.000    -1.532149   -1.146735
------------------------------------------------------------------------------
```

      i. Interpret the estimated coefficient assuming the relationship is causal.

      ii. Is the relationship quantitatively relevant? To answer this question, compare two groups of countries. Group 1 countries had a democracy score of 1 and an average log per capita income of 9.57 in the year 2000 (this includes quite a few countries including the EU member states). Group 2 countries had a democracy score of 0.5 and an average log per capita income of 7.85 (this includes countries like Albania, Burkina Faso, Kuwait, Paraguay, Turkey, and Ukraine). By how much could have the latter countries, according to the estimated model, closed the democracy gap by fully catching up economically?

[1] D. Acemoglu, S. Johnson, J. A. Robinson, P. Yared (2008), Income and Democracy, American Economic Review 98(3), 808-42.

iii. Discuss whether the relationship should be interpreted as being causal.

(b) A pooled OLS estimation of (1) yields (robust s.e.'s in brackets below the estimates)

$$\widehat{dem}_{it} = \underset{(0.035)}{0.706} dem_{i,t-1} + \underset{(0.010)}{0.072} inc_{i,t-1} + \hat{\mu}_t, \tag{2}$$

i. Discuss the pros and cons of adding a full set of time dummies.

ii. Why may it be sensible to include income with a lag?

iii. What is the rationale behind including the lagged democracy score as a regressor?

iv. Compute the short-term and long-term effects of an increase in income by 100 percent.

(c) A fixed effects estimation of (1) yields the results presented below.

```
. xtreg dem L.dem L.inc yr3-yr10 if sample==1, fe vce(robust)

Fixed-effects (within) regression               Number of obs      =       945
Group variable: code_numeric                    Number of groups   =       150

R-sq:  within  = 0.2417                         Obs per group: min =         1
       between = 0.8845                                        avg =       6.3
       overall = 0.6772                                        max =         9

                                                F(10,149)          =     16.06
corr(u_i, Xb)  = 0.7546                         Prob > F           =    0.0000

                          (Std. Err. adjusted for 150 clusters in code_numeric)
------------------------------------------------------------------------------
             |               Robust
         dem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dem |
          L1.|   .3786284   .0466924     8.11   0.000     .2863636    .4708931
             |
         inc |
          L1.|    .010415   .0316728     0.33   0.743    -.0521709    .0730009
             |
         yr3 |   -.044566   .0338314    -1.32   0.190    -.1114174    .0222853
         yr4 |  -.0744071   .0301114    -2.47   0.015    -.1339076   -.0149067
         yr5 |  -.1781914   .0311613    -5.72   0.000    -.2397666   -.1166163
         yr6 |   -.133589   .0286265    -4.67   0.000    -.1901554   -.0770226
         yr7 |  -.0731129   .0288599    -2.53   0.012    -.1301405   -.0160854
         yr8 |  -.0780685   .0253412    -3.08   0.002    -.1281431   -.0279939
         yr9 |  -.0432207   .0195065    -2.22   0.028    -.0817659   -.0046756
        yr10 |  -.0028151   .0190529    -0.15   0.883    -.0404639    .0348337
       _cons |   .3343984   .2696243     1.24   0.217    -.1983827    .8671796
-------------+----------------------------------------------------------------
     sigma_u |  .20460922
     sigma_e |  .18004117
         rho |   .5636116   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

i. Interpret the overall, within, and between R^2.

ii. Which of the POLS assumptions seems to be violated?

iii. Compute the short-term and long-term effects of an increase an increase in income by 100 percent.

iv. Give a potential explanation for why the estimation result differs so much from the POLS results (2).

v. Are you confident that the FE estimation results are valid?

4. Lundberg and Rose (2002) estimate the effect of the number of kids on fathers' labor supply and wage.[2] They consider the following two specifications:

$$y_{it} = \beta_1 MARR_{it} + \beta_2 NKID04_{it} + \beta_3 DKID5_{it} + \sum_j \beta_{age,j} DAGE_{j,it} + \sum_k \beta_{year,k} DYEAR_{k,it} + \sum_l \beta_{educ,l} DEDUC_{l,it} + c_i + u_{it} \quad (3)$$

and

$$y_{it} = \beta_1 MARR_{it} + \sum_{m=1}^{4} \beta_{nkid,m} DKID_{m,it} + \beta_3 DKID5_{it} + \sum_j \beta_{age,j} DAGE_{j,it} + \sum_k \beta_{year,k} DYEAR_{k,it} + \sum_l \beta_{educ,l} DEDUC_{l,it} + c_i + u_{it}, \quad (4)$$

where y_{it} is the outcome variable (either the log of the real hourly wage rate, or annual hours of work), $MARR_{it}$ is a marriage dummy (1=married), $NKID04_{it}$ is the number of kids if the man has four children or less and zero otherwise, $DKID5_{it}$ is a dummy variable for five or more children (1=at least five kids), and $DKID_{m,it}$ is a dummy variable indicating whether the man has exactly m kids. In addition, $DAGE_{j,it}$ is a series of dummy variables for each year of age of the individual, $DYEAR_{k,it}$ is a series of dummy variables representing the year of the observation, and $DEDUC_{l,it}$ is a series of dummy variables indicating the number of years of education.

(a) Why is it potentially important to control for age, year, and education effects? For each group of dummies give an example why leaving them out can lead to inconsistent estimates of the effect of the number of kids on fathers' labor supply or wage.

[2] Lundberg and Rose (2002), The Effects of Sons and Daughters on Men's Labor Supply and Wages, Review of Economics and Statistics 84(2), 251-268.

Tab. 1: Estimation results taken from Lundberg and Rose, 2002, p. 260

TABLE 5A.—THE EFFECT OF MARRIAGE AND CHILDREN ON ANNUAL HOURS WORKED (ENTIRE SAMPLE) (N = 26205)

	(1) OLS	(2) OLS	(3) OLS	(4) FE	(5) FE	(6) FE
Married	200.679	160.945	148.516	115.325	111.264	103.686
	(24.560)	(24.645)	(24.892)	(16.327)	(16.335)	(16.470)
Number of children (0 if none or >4)		45.86			38.416	
		(10.245)			(7.266)	
(Exactly) one child			68.297			82.023
			(22.983)			(14.849)
(Exactly) two children			138.562			108.165
			(25.595)			(17.729)
(Exactly) three children			138.922			113.230
			(34.375)			(24.544)
(Exactly) four children			126.268			152.212
			(66.625)			(36.551)
More than four children		−57.497	−34.916		38.074	49.624
		(133.137)	(132.643)		(62.147)	(62.319)
Two children − one child			*70.265*			*26.142*
			(24.215)			*(13.554)*
Three children − two children			*0.360*			*5.065*
			(30)			*(17.907)*
Four children − three children			*−12.654*			*38.982*
			(63.27)			*(31.111)*
R^2	0.04	0.04	0.04	0.45	0.45	0.45

Additional regressors include dummy variables for year of observation, years of education, and age. Standard errors in parentheses.

(b) What kinds of variables are captured in c_i? Give a few relevant examples.

(c) Explain how you would estimate this model taking the assumptions of the various estimators into account.

(d) Discuss the strict exogeneity assumption for the number of kids. Why may it fail?

(e) Table 1 reports some of their regression results using both pooled OLS and FE estimation.

i. Explain why the R^2 is (so much) higher for the FE estimator than for pooled OLS. Hint: Read Chapter 10.5.3 in the textbook about the least squares dummy variables (LSDV) estimator and conjecture that Lundberg and Rose used this estimator.

ii. (*) Prove that the LSDV estimator yields, in a standard error components model, the same estimator of $\boldsymbol{\beta}$ as the FE estimator. Hint: Have a look at the Frisch-Waugh-Lovell-Theorem at Wikipedia and show that the transformation matrix $\mathbf{M}_{\mathbf{X}_1}$ used there is identical to our $\mathbf{I}_N \otimes \mathbf{Q}_T$ matrix.

iii. Which of the estimation methods do you trust more?

iv. Now interpret the FE results. Compared to an unmarried man without kids, how many hours per year does a man work more

who (1) is married and has one kid, (2) is married an has four kids (3), is married and has six kids? Compare the results of the two specifications. Discuss.

Empirical exercises

1. Read Acemoglu et al. (2008) who analyze the effect of income on democracy.
   (a) Load their data set `AJRY_2008_data.dta`. Display important summary statistics (using the `xtsum` command). What do you learn?

   (b) Estimate the structural equation (1) by POLS, replicating column (1) of Table 1 of Acemoglu et al. (2008).

   (c) Compute the POLS-based long-run effect of income on democracy. Using the delta method to compute its standard error.

   (d) Estimate the structural equation (1) by FE, replicating column (2) of Table 1 of Acemoglu et al. (2008).

   (e) Estimate the structural equation (1) by RE and perform a Hausman test to choose between FE and RE.

   (f) Add the following controls: log of population, age structure (percent of the population in the age groups 0–15, 15–30, 30–45, and 45–60), and education (average years of schooling). Why may it be important to include these controls? Does it change the conclusions?

2. Read Cervellati et al. (2014) who reassess the findings of Acemoglu et al. (2008).[3] Their baseline specification is

$$dem_{it} = \beta_1 dem_{i,t-1} + \beta_2 inc_{i,t-1} + \beta_3 (inc_{i,t-1} \cdot f_i) + \mu_t + c_i + u_{it}, \quad (5)$$

   where f_i indicates whether a country is a former European colony or not.

   (a) Explain why the effect of income on democracy may differ between former colonies and other countries.

[3] M. Cervellati, F. Jung, U. Sunde, and T. Vischer (2014), Income and Democracy: Comment, American Economic Review 104(2), 707-19.

(b) Load the data set `Cervellati_2014_data.dta` and construct a scatter plot of lagged income against the democracy score both for all countries and separated between former colonies and other countries.

(c) Construct a scatter plot of lagged income, net of past democracy, year fixed effects and country fixed effects, against the democracy score both for all countries and separated between former colonies and other countries. This replicates panels (b) and (c) of Figure 1 of Cervellati et al. (2014).

(d) Estimate (5) by FE, replicating column (2) of Table 1 of Cervellati et al. (2014). What are the short-term and long-term effects of income on democracy for (i) former colonies and (ii) other countries? Are the differences between country groups economically relevant and statistically significant?