AFRICAN ECONOMIC RESEARCH CONSORTIUM

Collaborative Masters Programme in Economics for Anglophone Africa
(Except Nigeria)

JOINT FACILITY FOR ELECTIVES (JFE) 2011
JUNE - SEPTEMBER

ECONOMETRICS THEORY AND PRACTICE II

DR. MOSES SICHEI*

Date:6th September, 2011

LECTURE 7B: PANEL DATA ECONOMETRICS I

Objectives:

The main objective of the lecture is to provide motivation for panel data models. Specifically, the lecture presents the following:

- Places panel data in the context of other data types
- Types of panel data types
- Advantages of panel data
- Limitations of panel data
- Overview of panel data models

Key words

* KSMS Research Centre
Mobile:+254 723383505;Email: sichei@yahoo.co.uk or Sicheimm@centralbank.go.ke or sicheimm@ksms.or.ke

- Fixed effects model(FEM)
- Generalised least squares
- Least squares dummy variable (LSDV)

- Balanced panel data
- Between estimator
- Cross-section oriented panel data
- Dynamic panel
- Macro-panel data
- Micro-panel data
- Nonstationary panel data
- One-way error components model
- Panel data
- Pooled data
- Pseudo-panel
- Random effects model (REM)
- Rotating panels
- Seemingly unrelated regression model
- Spatial panel data
- Static panel data
- Stationary panel data
- Synthetic panel data
- Unbalanced panel data
- Within estimation
-

1. INTRODUCTION AND MOTIVATION

1.1 Types of Data

1.1.1 Cross-section data

- Values of one or more variables are collected for several sample units/economic entities at the same point in time.
- In other words it is a **snapsho**t at a point in time
- Examples
    - Poverty rates in different countries in Africa at a particular point in time
    - Econometrics marks for the 2011 CMAP group

- Household survey data for Uganda

- Cross-section models are predominantly equilibrium models that generally do not shed light on intertemporal dependence of events
- They fail to resolve fundamental issues about the sources of persistent behaviour
- For instance what's the main cause of high non-performing loans in country A?

1.1.2 Time series data

- Observe the values of one or more variables over time e.g. GDP, money supply for several years.
- It is like a movie
- They shed light on intertemporal dependence of events
- E.g. autoregressive distributed lag models, error correction models etc.

1.1.3 Panel data (cross-section and time series)

- Cross-section repeatedly sampled over time but where the same economic agent has been followed throughout the period of the sample.
- An example is the average marks for the CMAP econometrics course for each university over the period 2001-2011.

Period	University of Nairobi	University of Dar Es Salaam	Makerere University	University of Ghana	University of Addis Ababa	University of Malawi	University of Zimbabwe	University of Botswana	University of Cape Coast	University of Mauritius	University of Namibia
2001											
2002											
2003											
2004											
2005											

2006											
2007											
2008											
2009											
2010											
2011											

- In other words panel data combines cross-section ("**picture or snapshot", or space**) with time series (**"path."**, **movie**)

Other terms used;

- **Pooled data** (pooling of time series and cross-section observations)
- **Micro-panel data**
- **Combination of time series and cross-section data**
- **Micro-panel data**
- **Longitudinal data** (Study over time of a variable or group of subjects).
- **Event history** (study of the movement over time of subjects through successive states or conditions)
- **Cohort analysis** (e.g. following the career path of the first CMAP graduates)

⇒ All these terminologies essentially connote movements over time of cross-sectional units.

⇒Thus panel data is used as a generic term to include one or more of these situations.

⇒Regressions based on such data are called panel data regression models

Examples

- Gravity model of trade, where you observe trade figures for different countries/products over time
- Investment model, where your cross-sections are the firms observed over time
- Studies dealing with a panel of commercial banks
- Etc.

1.2 Structures of Panel Data

(a) **Cross-section oriented panel data**. The number of cross-sections (N) is more than the time dimension (T) .e.g. study covering 24 banks over 10 years. This is the original panel data

(b) **Time-series oriented panel data**. The time dimension (T) is greater than the cross-sections (N) e.g. Study of the demand for 4 different oil products covering a period of say 10 years. This is quite common in macroeconomics

(c) **Balanced panel data**. This is panel data where there is no missing observations for every cross-section

(d) **Unbalanced panel data**. This is the case, where the cross-sections do not have the same number of data observations. In other words some cross-sections do not have data. For example when studying Ghana's trade data to a number of countries in Africa including South Africa, There would be no exports figures before 1994 due to sanctions imposed on South Africa.

(e) **Rotating panels.** This is a case where in order to keep the same number of economic agents in a survey; the fraction of economic agents that drops from the sample in the second period is replaced by an equal number of similar economic agents that are freshly surveyed. This is a necessity in survey panels where the same economic agent (say household) may not want to be interviewed again and again.

(f) **Pseudo-Panels/synthetic panels.** This panel data that is close to a genuine panel data structure. For instance for some countries, panel data may not exist. Instead the researcher may find annual household survey based on a large random sample of the population. For instance in Kenya there are household surveys for 1993, 1994, 1997 and the recent KHIBS 2006. For these repeated cross-section surveys, it may be impossible to track the same household over time as required in a genuine panel. In Pseudo panels **cohorts** are tracked (e.g. males borne between 1970 to 1980). For large samples, successive surveys will generate random samples of members of each cohort. We can then estimate economic relations based on means rather than individual observations.

(g) **Spatial Panels.** This is panel data dealing with space. For instance cross-section of countries, regions, states. These aggregate units are likely to exhibit cross-sectional correlation that has to be dealt with using special methods (spatial econometrics)

(h) Limited dependent/nonlinear panel data. This is panel data where the dependent variable is not completely continuous-binary(logit/probit models), hierarchical (nested logit models), ordinal (ordered logit/probit), categorical(multinomial logit/probit), count models(poisson and negative binomial), truncated (truncated regression), censored (tobit), sample selection(Heckit model)

1.3 Types of Panel Data Models

(a) Static Panel data Models vs Dynamic Panel Data Model. Static panel data model has no lagged dependent variable on the rhs.

(c) Stationary Panel Data Model vs Non Stationary Panel Data Model. Stationary panel data model contain stationary variables (i.e. I(0) variables) as opposed to non-stationary variables

1.4 Benefits of Panel Data?

1. **More informative data, more variability, less collinearity amongst variables, more degress of freedom and more efficiency;**
   - Time series studies are faced with multi-collinearity in most cases for instance in studying say the demand for beer in Kenya using time series, there is likely to be high collinearity between price and income in aggregate time series data. This is less likely with a panel across the 8 provinces in Kenya since the cross-section dimension adds a lot of variability, adding more informative data on price and income. The idea is that the variation in the data can be decomposed into variation between the 8 provinces of different sizes and characteristics and variation within each prince over time.
   - Blows up degrees of freedom (NxT);
   - Increased precision in estimation (more efficiency). With additional more informative data, we can produce more reliable parameter estimates.

2. **Panel data are better able to study dynamics of adjustment**. Panel data are better suited for studying the duration of economic states like unemployment and poverty and if such panels are long enough, they can shed some light on the speed of adjustments of to economic policy changes.

- E.g. the effects of free primary education on poverty.
- Questions such as determining whether families' experiences of poverty, unemployment and dependency ratios are transitory or chronic necessitate the use of panels. By studying the repeated cross-section of observations, panel data are better suited to study the dynamics of change.
- Estimation of intertemporal relations, life cycles and inter-generational models

3. Idenfitication and discrimination between competing hypotheses

- Panel data provides sequential observations for a number of individuals and thus allow us to distinguish inter-individual differences from intra-individual differences and to construct proper recursive structure for studying the issue in question through a before-and- after effect.
- For instance does union membership in Kenya increase or decrease wages? We need to observe a worker moving from union to nonunion jobs or vice versa. Holding the individual's characteristics constant, we will be better equipped to determine whether union membership affects wage and by how much.

4. Provides micro foundations for aggregate data analysis

- Aggregate data analysis often invokes the "representative agent" assumption
- However, if microunits are heterogeneous, policy evaluation based on aggregate data can be grossly misleading
- Additionally, the prediction of aggregate outcomes using aggregate data can be less accurate than the predictions based on micro-equations
- Panel data are usually gathered on micro units like individuals, firms, households, countries etc.
- Many variables can be more accurately measured at the micro level and biases resulting from aggregation over firms or individuals are eliminated

5. Reduces estimation bias

- Ommitted variable bias

- Bias induced by the dynamic structure of the model
- Simultaneity bias
- Measurement errors

1.5 Limitations of Panel Data Analysis

1. Design and data collection problems

- Coverage (incomplete account of the population of interest)
- Non-response (due to lack of cooperation of the respondent-fear for use of the results(tax?) or because of interviewer error
- Recall (respondents not remembering correctly)
- Freq of interviews
- Time in sample bias-is observed when a significantly different level for a characteristic occurs in the first interview than in later interviews, when ideally one would expect the same level

2. Distortions of measurement error

Measurement errors may arise because of

- Faulty responses due to unclear questions, memory errors, deliberate distortions (e.g.prestige bias); inappropriate informants, misrecording of responses

3. Selectivity problems

These include;

- *Self-selectivity*-For instance people choose not to work because of the reservation wage is higher than the offered wage. In this case we only observe the characteristics of these individuals but not their wage. Panel data does not solve such a problem
- *Non/partial-responses*-Occurs mainly at the initial wave of the panel due to refusal to participate, nobody at home, untraced sample unit, etc. This cause loss in efficiency as well as serious identification problems for the population parameters
- Attrition-Respondents may die, or move or find that the cost of responding is too high. In order to counter the effects of attrition, **rotating panels** are sometimes used,

where a fixed percentage of the respondents are replaced in every wave to replenish the sample.

4. **Very short time series dimension**
   - Sometimes data for the problem at hand has a short time span for each individual (micropanels).
   - This means that asymptotic refinements which rely crucially on the number of individuals tending to infinity may not be useful.
5. **Cross-section dependence**
   - Macropanels on countries or regions may lead to misleading inference.
   - Most analyses assume independence
   - When we have dependence between cross-sections, it becomes complicated
   - More on this when we handle panel unit roots and panel cointegration

2. OVERVIEW OF PANEL DATA MODELS

- Panel data model notation differs from a regular time series or cross-section regression in that it has a double subscript on its variables ; y_{it}, x_{it} ;

2.1 Panel Data Models

- A very general linear model for panel data permits the intercept and slope coefficients to vary over both individual and time

$$y_{it} = \alpha_{it} + x'_{it}\beta_{it} + \varepsilon_{it}\ ,\ i = 1,...,N\ , t = 1,...,T \qquad (1)$$

- Where y_{it} is a scalar dependent variable, x_{it} is a $k \times 1$ vector of independent variables, ε_{it} is a scalar disturbance term, i indexes individuals (firms, country etc.), t indexes time.
- This model is too general and **not estimable** as there are more parameters to estimate than observations
- **More restrictions** need to be placed on the extent to which α_{it} and β_{it} can vary with i and t and on the behaviour of the error term ε_{it}

2.2 The Pooled Data Model

- This is the most restrictive model that specifies constant coefficients, which is the usual assumption about cross-section analysis

$$y_{it} = \alpha + x'_{it}\beta + \varepsilon_{it} \qquad (2)$$

Where i denotes households, individuals, firms, countries etc and t denotes time.

- We assume that errors are homoscedastic and serially independent both ***within*** and ***between*** individuals(cross-sections).

$Var(\varepsilon_{it}) = \sigma^2$

$Cor(\varepsilon_{it}, \varepsilon_{js}) = 0$ when $i \neq j$ and/or $t \neq s$

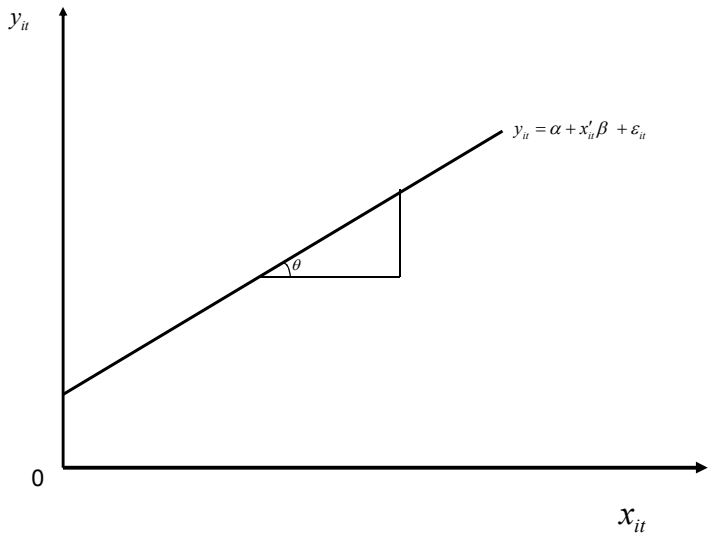
- The **marginal effects** β of the set of k vector of time-varying characteristics x_{it} are taken to be **common across** *i* and *t*, although this assumption can itself be tested.
- If the model is correctly specified and regressors are uncorrelated with the error term, the pooled OLS will product consistent and efficient estimates for the parameters
- This is the pooled least squares model.

$$\hat{\beta} = \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}'_{it}\tilde{y}_{it}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}'_{it}\tilde{x}_{it}}$$

$$\text{Where } \bar{x} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}\,,\ \bar{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}\,,\ \tilde{x}_{it} = x_{it} - \bar{x}\,,\ \tilde{y}_{it} = y_{it} - \bar{y} \qquad (3)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

- This formulation does **not distinguish** between **two different individuals** and the same individual at **two different points** in time
- This feature undermines the accuracy of the approach when differences do exist between cross-sectional units.
- Nonetheless, the increase in the sample by pooling data across time generates an improvement in efficiency relative to a single cross-section.



- Here we do not use any panel information. The data are treated as if there was only one single index.

2.3 Traditional Panel Data Model

- In this case the constant term, α_i, varies from individual to individual.

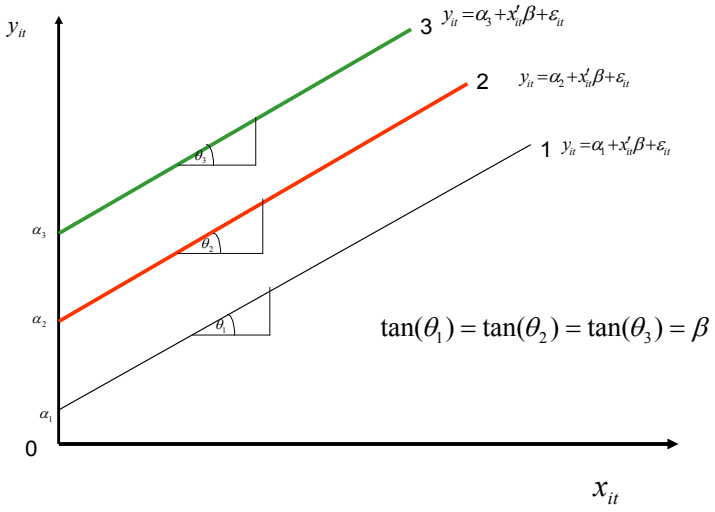
$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it} \quad (4)$$

- We assume that errors are homoscedastic and serially independent both ***within*** and ***between*** individuals(cross-sections).

$Var(\varepsilon_{it}) = \sigma^2$

$Cor(\varepsilon_{it}, \varepsilon_{js}) = 0$ when $i \neq j$ and/or $t \neq s$

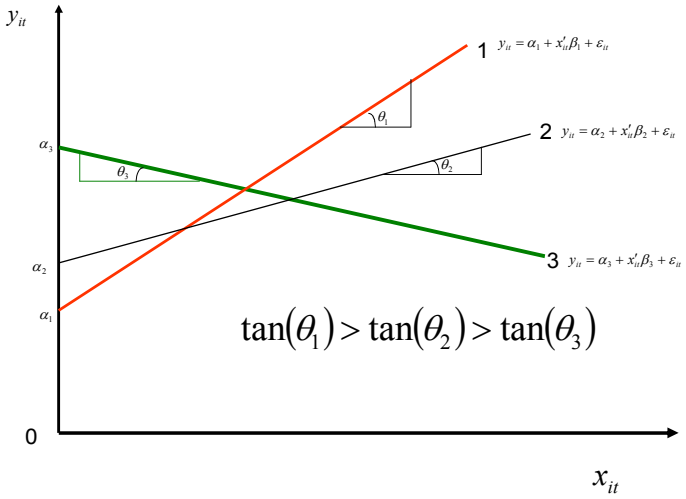
- This is what we refer to in panel parlance as **individual (unobserved) heterogeneity**.
- The slopes are the same for all individuals i.e.
- Its graphical form is as follows;



2.4 Traditional Seemingly Unrelated Regression (SUR) Model

- The constant terms, α_i, and slope coefficients, β_i,vary from individual to individual.

$$y_{it} = \alpha_i + x'_{it}\beta_i + \varepsilon_{it}$$



- In the SUR models, the error terms are assumed to be contemporaneously correlated and heteroscedastic ***between*** individuals.

$$Var(\varepsilon_{it}) = \sigma_i^2$$

$Cor(\varepsilon_{it}, \varepsilon_{jt}) = \sigma_{ij}$ Contemporaneous (same period) correlation

$Cor(\varepsilon_{it}, \varepsilon_{js}) = 0$ when $t \neq s$

2.5 Which model is appropriate for my Data?

- ❖ Large number of independent individuals observed for a few time periods (N>>T). This is common in cross-sectional panels.
  - o It is not possible to estimate different individual slopes, β_i, for all the exogenous variables. The **panel data model** is the most appropriate.
- ❖ There are medium length time series for relatively few individuals (say countries, firms, sectors, banks,etc.). T>N.
  - o In this case the SUR model may be appropriate.
  - o Efficient SUR estimation is mainly used when $T \geq N$.
  - o Equation by Equation OLS is used if $K \leq T \leq N$

- ❖ In terms of unrestrictiveness, the relationship is as follows:

Pooled (i.e.most restrictive**)<Panel<SUR(** i.e.most unrestrictive**)**

3. **One-Way Error Component Model**

- This model allows cross-section heterogeneity in the error term.
- From the traditional panel data model

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it} \quad (5)$$

- The error term in equation 5 is decomposed into;

$$\varepsilon_{it} = \mu_i + \nu_{it} \quad (6)$$

- Where μ_i denotes the *unobservable* individual specific effect and ν_{it} denotes idiosyncratic errors or idiosyncratic disturbances, which change across time and cross-section.

- μ_i is time invariant (same for all the time) and accounts for any individual-specific effect that is not included in the regression.

3.1 Meaning of the unobservable individual specific effects μ_i

- These refer to unobservable individual specific effects which are not included in the equation because of :
    - We do not know exactly how to specify them explicitly
    - We know but have no data
- We simply want to acknowledge their existence
- For instance in a production function utilizing data on firms across time, μ_i refers to the unobservable entrepreneurial or management skills of the firm executives

The other terminologies given to μ_i are;

- Latent variable
- Unobserved heterogeneity

- Substituting Equation 6 in 5 yields the following one-way error component model;

$$y_{it} = \alpha + \mu_i + x'_{it}\beta + \nu_{it} \quad (7)$$

- Note here that $\alpha = \frac{1}{N}\sum_{i=1}^{N}\alpha_i$

$$\mu_i = \alpha_i - \alpha$$

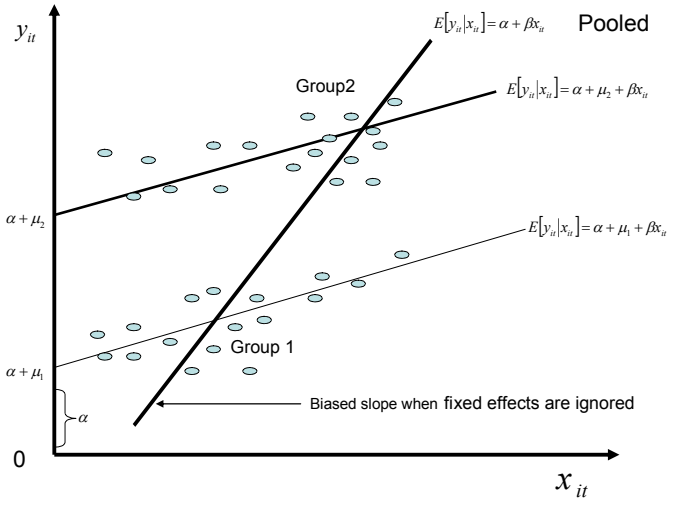
α is the average individual effect while μ_i is the individual deviation from the average (recall the reference class is multinomial logit model)

4 Fixed Effects Model (FEM)

- This is appropriate when differences between individual economic agents may reasonably be viewed as parametric shifts in the regression function itself.
- Suppose we have a simple linear panel regression model of the form ;

$$y_{it} = \alpha + \mu_i + x'_{it}\beta + \nu_{it} \quad (8)$$

- The $\alpha + \mu_i$ are possibly correlated with the regressors x_{it}
- The following figure shows how the FEM handles the heterogeneity issue.



4.1 Estimation of FEM

- The challenge of estimation is the presence of the N individual-specific effects that increase as $N \to \infty$
- Nonetheless, there are several methods that can be applied
    - Least squares dummy variables (LSDV) which is a direct OLS with indicator (dummy variables) for each of the N fixed effects
    - Use OLS in the WITHIN estimation context
    - Generalized Leased squares in the WITHIN model context
    - Maximum likelihood estimation conditional on the individual means $\bar{y}_i, i = 1,2,...,N$
    - OLS in first differences

4.1.1 Least Squares Dummy Variable (LSDV)

- This approach assumes that any difference across economic agents can be captured by shifts in the intercepts of a standard OLS regression.

$$y_{it} = \alpha + \mu_i + x'_{it}\beta + \nu_{it}$$

- We estimate an LSDV model first by defining a series of individual-specific dummies variables.

- In principle one simply estimates the OLS regression of y_{it} on x_{it} and a set of N-1 indicator variables $d_{1t,}d_{2it},...d_{(N-1)t}$
- The resulting estimator of β turns out to equal the within estimator (running a regression through the mean)
- This is a special case of the Frisch-Waugh-Lovell theorem. You have been using this theorem: running a regression through the origin (after subtracting the mean) produces the same slope coefficients as running it with an intercept).
- The theorem was introduced by **Frisch and Waugh (1933),** and then reintroduced by Lovell (1963).
- Read pages 62-75 of Econometric theory and methods by Davidson and Mackinnon (2004) for more details of the theorem

Digression: Different ways of stacking data

- Suppose we studying private consumption in 12 Africa countries over the period 1998-2003
- We have data on real consumption and income for the different African countries

Dependent variable: consumption

(i) stacked vertically to create $NT \times 1$ vector

Country	Period	Consumption
Botswana	1998	7180.3
Botswana	1999	7533.5
Botswana	2000	7841.1
Botswana	2001	7919.2
Botswana	2002	8085.2
Botswana	2003	8222.9
Burkina Faso	1998	1283027.4
Burkina Faso	1999	1297642.2
Burkina Faso	2000	1306400.0
Burkina Faso	2001	1411715.4
Burkina Faso	2002	1513999.2
Burkina Faso	2003	1626124.7
Burundi	1998	462066.7
Burundi	1999	492055.5
Burundi	2000	465738.0
Burundi	2001	469289.9
Burundi	2002	491701.3
Burundi	2003	486195.2
Kenya	1998	596883.1
Kenya	1999	594332.1
Kenya	2000	609862.0
Kenya	2001	629103.7
Kenya	2002	650968.4
Kenya	2003	680065.0
Madagascar	1998	21830.2
Madagascar	1999	22441.8
Madagascar	2000	22483.0
Madagascar	2001	22443.8
Madagascar	2002	21150.2
Madagascar	2003	22985.4
Mauritius	1998	69552.9
Mauritius	1999	71594.9
Mauritius	2000	73939.3
Mauritius	2001	76048.7
Mauritius	2002	78570.9
Mauritius	2003	82602.2
Morocco	1998	240.3
Morocco	1999	233.4
Morocco	2000	243.0
Morocco	2001	256.4
Morocco	2002	256.1
Morocco	2003	261.7
Nigeria	1998	3307.9
Nigeria	1999	2255.7
Nigeria	2000	2446.5
Nigeria	2001	3068.0
Nigeria	2002	3665.8
Nigeria	2003	3424.9
Rwanda	1998	588.0
Rwanda	1999	595.6
Rwanda	2000	641.9
Rwanda	2001	676.1
Rwanda	2002	740.4
Rwanda	2003	769.9
Sierra Leone	1998	1180237.6
Sierra Leone	1999	1032168.8
Sierra Leone	2000	1142680.0
Sierra Leone	2001	1369830.5
Sierra Leone	2002	1547871.3
Sierra Leone	2003	1613277.6
South Africa	1998	516925.9
South Africa	1999	531213.0
South Africa	2000	556652.0
South Africa	2001	579316.4
South Africa	2002	598804.9
South Africa	2003	614082.8
Tanzania	1998	5610.4
Tanzania	1999	6003.2
Tanzania	2000	6069.6
Tanzania	2001	6579.9
Tanzania	2002	7064.1
Tanzania	2003	7974.2

- This is how data for stata, SAS and PCGIVE should be organized
- Data should be organized this way for Eviews if you would like to use dynamic panel methods

(ii) Horizontal to create $NT \times N$ matrix

Country	Period	rcons_Bots	rcons_BurkF	rcons_Bur	rcons_Ken	rcons_Madag	rcons_Maurit	rcons_Mor	rcons_Nig	rcons_Rwa	rcons_SierL	rcons_rsa	rcons_Tan
Botswana	1998	7180.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Botswana	1999	7533.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Botswana	2000	7841.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Botswana	2001	7919.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Botswana	2002	8085.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Botswana	2003	8222.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burkina Faso	1998	0.0	1283027.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burkina Faso	1999	0.0	1297642.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burkina Faso	2000	0.0	1306400.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burkina Faso	2001	0.0	1411715.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burkina Faso	2002	0.0	1513999.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burkina Faso	2003	0.0	1626124.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burundi	1998	0.0	0.0	462066.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burundi	1999	0.0	0.0	492055.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burundi	2000	0.0	0.0	465738.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burundi	2001	0.0	0.0	469289.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burundi	2002	0.0	0.0	491701.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Burundi	2003	0.0	0.0	486195.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kenya	1998	0.0	0.0	0.0	596883.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kenya	1999	0.0	0.0	0.0	594332.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kenya	2000	0.0	0.0	0.0	609862.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kenya	2001	0.0	0.0	0.0	629103.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kenya	2002	0.0	0.0	0.0	650968.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kenya	2003	0.0	0.0	0.0	680065.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Madagascar	1998	0.0	0.0	0.0	0.0	21830.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Madagascar	1999	0.0	0.0	0.0	0.0	22441.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Madagascar	2000	0.0	0.0	0.0	0.0	22483.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Madagascar	2001	0.0	0.0	0.0	0.0	22443.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Madagascar	2002	0.0	0.0	0.0	0.0	21150.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Madagascar	2003	0.0	0.0	0.0	0.0	22985.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mauritius	1998	0.0	0.0	0.0	0.0	0.0	69552.9	0.0	0.0	0.0	0.0	0.0	0.0
Mauritius	1999	0.0	0.0	0.0	0.0	0.0	71594.9	0.0	0.0	0.0	0.0	0.0	0.0
Mauritius	2000	0.0	0.0	0.0	0.0	0.0	73939.3	0.0	0.0	0.0	0.0	0.0	0.0
Mauritius	2001	0.0	0.0	0.0	0.0	0.0	76048.7	0.0	0.0	0.0	0.0	0.0	0.0
Mauritius	2002	0.0	0.0	0.0	0.0	0.0	78570.9	0.0	0.0	0.0	0.0	0.0	0.0
Mauritius	2003	0.0	0.0	0.0	0.0	0.0	82602.2	0.0	0.0	0.0	0.0	0.0	0.0
Morocco	1998	0.0	0.0	0.0	0.0	0.0	0.0	240.3	0.0	0.0	0.0	0.0	0.0
Morocco	1999	0.0	0.0	0.0	0.0	0.0	0.0	233.4	0.0	0.0	0.0	0.0	0.0
Morocco	2000	0.0	0.0	0.0	0.0	0.0	0.0	243.0	0.0	0.0	0.0	0.0	0.0
Morocco	2001	0.0	0.0	0.0	0.0	0.0	0.0	256.4	0.0	0.0	0.0	0.0	0.0
Morocco	2002	0.0	0.0	0.0	0.0	0.0	0.0	256.1	0.0	0.0	0.0	0.0	0.0
Morocco	2003	0.0	0.0	0.0	0.0	0.0	0.0	261.7	0.0	0.0	0.0	0.0	0.0
Nigeria	1998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3307.9	0.0	0.0	0.0	0.0
Nigeria	1999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2255.7	0.0	0.0	0.0	0.0
Nigeria	2000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2446.5	0.0	0.0	0.0	0.0
Nigeria	2001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3068.0	0.0	0.0	0.0	0.0
Nigeria	2002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3665.8	0.0	0.0	0.0	0.0
Nigeria	2003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3424.9	0.0	0.0	0.0	0.0
Rwanda	1998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	588.0	0.0	0.0	0.0
Rwanda	1999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	595.6	0.0	0.0	0.0
Rwanda	2000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	641.9	0.0	0.0	0.0
Rwanda	2001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	676.1	0.0	0.0	0.0
Rwanda	2002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	740.4	0.0	0.0	0.0
Rwanda	2003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	769.9	0.0	0.0	0.0
Sierra Leone	1998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1180237.6	0.0	0.0
Sierra Leone	1999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1032168.8	0.0	0.0
Sierra Leone	2000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1142680.0	0.0	0.0
Sierra Leone	2001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1369830.5	0.0	0.0
Sierra Leone	2002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1547871.3	0.0	0.0
Sierra Leone	2003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1613277.6	0.0	0.0
South Africa	1998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	516925.9	0.0
South Africa	1999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	531213.0	0.0
South Africa	2000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	556652.0	0.0
South Africa	2001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	579316.4	0.0
South Africa	2002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	598804.9	0.0
South Africa	2003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	614082.8	0.0
Tanzania	1998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5610.4
Tanzania	1999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6003.2
Tanzania	2000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6069.6
Tanzania	2001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6579.9
Tanzania	2002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7064.1
Tanzania	2003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7974.2

The following stacking is used by Eviews software when you are not interested in using dynamic panel or the earlier version of Eviews software cannot allow you to do so (e.g.Eviews 3.1)

Period	rcons_Bots	rcons_BurkF	rcons_Bur	rcons_Ken	rcons_Madag	rcons_Maurit	rcons_Mor	rcons_Nig	rcons_Rwa	rcons_SierL	rcons_rsa	rcons_Tan
1990	5587.5	944244.1	652316.1	534498.2	18388.6	48387.5	203.9	2003.8	680.7	1351781.6	416324.7	4179.1
1991	6361.7	922686.9	616476.2	521072.4	19200.5	50061.0	230.1	2538.2	639.0	1538679.5	417907.8	4188.8
1992	6517.6	954058.8	686615.5	513099.7	18385.7	52803.7	219.0	3192.2	646.7	1359474.4	419952.9	4391.4
1993	5989.5	978916.2	712774.9	414525.6	19496.6	55480.3	208.9	2700.9	624.1	1485785.2	430586.1	4471.0
1994	6344.1	878232.0	733701.2	382161.6	19770.5	58196.4	229.5	2221.0	546.1	1403960.5	447390.1	4490.5
1995	6361.3	998737.2	548712.2	485436.4	19840.0	60635.0	221.2	2857.5	440.4	1445824.7	473595.4	4585.6
1996	6397.7	1103507.1	411135.4	496801.0	19766.0	63251.7	242.1	3391.5	494.8	1470327.2	494634.1	4684.2
1997	6633.4	1112742.4	409956.4	562446.2	21313.3	65509.1	230.5	3222.4	579.4	1322968.4	510869.8	5115.2
1998	7180.3	1283027.4	462066.7	596883.1	21830.2	69552.9	240.3	3307.9	588.0	1180237.6	516925.9	5610.4
1999	7533.5	1297642.2	492055.5	594332.1	22441.8	71594.9	233.4	2255.7	595.6	1032168.8	531213.0	6003.2
2000	7841.1	1306400.0	465738.0	609862.0	22483.0	73939.3	243.0	2446.5	641.9	1142680.0	556652.0	6069.6
2001	7919.2	1411715.4	469289.9	629103.7	22443.8	76048.7	256.4	3068.0	676.1	1369830.5	579316.4	6579.9
2002	8085.2	1513999.2	491701.3	650968.4	21150.2	78570.9	256.1	3665.8	740.4	1547871.3	598804.9	7064.1
2003	8222.9	1626124.7	486195.2	680065.0	22985.4	82602.2	261.7	3424.9	769.9	1613277.6	614082.8	7974.2

Note that if the above data is treated as a matrix, you can simply stack it vertically by using the vectorisation algebra (available as option in matrix algebra in most software like stata and eviews)

$$A=\begin{bmatrix}1 & 3\\ 4 & 7\end{bmatrix} \qquad Vec(A)=\begin{bmatrix}1\\ 4\\ 3\\ 7\end{bmatrix}$$

Independent variable: income

- The same stacking as the dependent variable can be done depending on the software

Individual specific dummy variables for LSDV model: Creates $NT \times N$ matrix

Country	Period	Bots	BurkF	Bur	Ken	Madag	Maurit	Mor	Nig	Rwa	SierL	rsa	Tan	
Botswana	1998		1	0	0	0	0	0	0	0	0	0	0	0
Botswana	1999		1	0	0	0	0	0	0	0	0	0	0	0
Botswana	2000		1	0	0	0	0	0	0	0	0	0	0	0
Botswana	2001		1	0	0	0	0	0	0	0	0	0	0	0
Botswana	2002		1	0	0	0	0	0	0	0	0	0	0	0
Botswana	2003		1	0	0	0	0	0	0	0	0	0	0	0
Burkina Faso	1998		0	1	0	0	0	0	0	0	0	0	0	0
Burkina Faso	1999		0	1	0	0	0	0	0	0	0	0	0	0
Burkina Faso	2000		0	1	0	0	0	0	0	0	0	0	0	0
Burkina Faso	2001		0	1	0	0	0	0	0	0	0	0	0	0
Burkina Faso	2002		0	1	0	0	0	0	0	0	0	0	0	0
Burkina Faso	2003		0	1	0	0	0	0	0	0	0	0	0	0
Burundi	1998		0	0	1	0	0	0	0	0	0	0	0	0
Burundi	1999		0	0	1	0	0	0	0	0	0	0	0	0
Burundi	2000		0	0	1	0	0	0	0	0	0	0	0	0
Burundi	2001		0	0	1	0	0	0	0	0	0	0	0	0
Burundi	2002		0	0	1	0	0	0	0	0	0	0	0	0
Burundi	2003		0	0	1	0	0	0	0	0	0	0	0	0
Kenya	1998		0	0	0	1	0	0	0	0	0	0	0	0
Kenya	1999		0	0	0	1	0	0	0	0	0	0	0	0
Kenya	2000		0	0	0	1	0	0	0	0	0	0	0	0
Kenya	2001		0	0	0	1	0	0	0	0	0	0	0	0
Kenya	2002		0	0	0	1	0	0	0	0	0	0	0	0
Kenya	2003		0	0	0	1	0	0	0	0	0	0	0	0
Madagascar	1998		0	0	0	0	1	0	0	0	0	0	0	0
Madagascar	1999		0	0	0	0	1	0	0	0	0	0	0	0
Madagascar	2000		0	0	0	0	1	0	0	0	0	0	0	0
Madagascar	2001		0	0	0	0	1	0	0	0	0	0	0	0
Madagascar	2002		0	0	0	0	1	0	0	0	0	0	0	0
Madagascar	2003		0	0	0	0	1	0	0	0	0	0	0	0
Mauritius	1998		0	0	0	0	0	1	0	0	0	0	0	0
Mauritius	1999		0	0	0	0	0	1	0	0	0	0	0	0
Mauritius	2000		0	0	0	0	0	1	0	0	0	0	0	0
Mauritius	2001		0	0	0	0	0	1	0	0	0	0	0	0
Mauritius	2002		0	0	0	0	0	1	0	0	0	0	0	0
Mauritius	2003		0	0	0	0	0	1	0	0	0	0	0	0
Morocco	1998		0	0	0	0	0	0	1	0	0	0	0	0
Morocco	1999		0	0	0	0	0	0	1	0	0	0	0	0
Morocco	2000		0	0	0	0	0	0	1	0	0	0	0	0
Morocco	2001		0	0	0	0	0	0	1	0	0	0	0	0
Morocco	2002		0	0	0	0	0	0	1	0	0	0	0	0
Morocco	2003		0	0	0	0	0	0	1	0	0	0	0	0
Nigeria	1998		0	0	0	0	0	0	0	1	0	0	0	0
Nigeria	1999		0	0	0	0	0	0	0	1	0	0	0	0
Nigeria	2000		0	0	0	0	0	0	0	1	0	0	0	0
Nigeria	2001		0	0	0	0	0	0	0	1	0	0	0	0
Nigeria	2002		0	0	0	0	0	0	0	1	0	0	0	0
Nigeria	2003		0	0	0	0	0	0	0	1	0	0	0	0
Rwanda	1998		0	0	0	0	0	0	0	0	1	0	0	0
Rwanda	1999		0	0	0	0	0	0	0	0	1	0	0	0
Rwanda	2000		0	0	0	0	0	0	0	0	1	0	0	0
Rwanda	2001		0	0	0	0	0	0	0	0	1	0	0	0
Rwanda	2002		0	0	0	0	0	0	0	0	1	0	0	0
Rwanda	2003		0	0	0	0	0	0	0	0	1	0	0	0
Sierra Leone	1998		0	0	0	0	0	0	0	0	0	1	0	0
Sierra Leone	1999		0	0	0	0	0	0	0	0	0	1	0	0
Sierra Leone	2000		0	0	0	0	0	0	0	0	0	1	0	0
Sierra Leone	2001		0	0	0	0	0	0	0	0	0	1	0	0
Sierra Leone	2002		0	0	0	0	0	0	0	0	0	1	0	0
Sierra Leone	2003		0	0	0	0	0	0	0	0	0	1	0	0
South Africa	1998		0	0	0	0	0	0	0	0	0	0	1	0
South Africa	1999		0	0	0	0	0	0	0	0	0	0	1	0
South Africa	2000		0	0	0	0	0	0	0	0	0	0	1	0
South Africa	2001		0	0	0	0	0	0	0	0	0	0	1	0
South Africa	2002		0	0	0	0	0	0	0	0	0	0	1	0
South Africa	2003		0	0	0	0	0	0	0	0	0	0	1	0
Tanzania	1998		0	0	0	0	0	0	0	0	0	0	0	0
Tanzania	1999		0	0	0	0	0	0	0	0	0	0	0	0
Tanzania	2000		0	0	0	0	0	0	0	0	0	0	0	0
Tanzania	2001		0	0	0	0	0	0	0	0	0	0	0	0
Tanzania	2002		0	0	0	0	0	0	0	0	0	0	0	0
Tanzania	2003		0	0	0	0	0	0	0	0	0	0	0	0

- Note the fact that the last cross-section is not coded with 1 to avoid the problem of dummy variable trap (i.e. perfect multi-collinearity)

Use of Kronecker product

Recall that $A \otimes B = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1\cdot 0 & 1\cdot 3 & 2\cdot 0 & 2\cdot 3 \\ 1\cdot 2 & 1\cdot 1 & 2\cdot 2 & 2\cdot 1 \\ 3\cdot 0 & 3\cdot 3 & 1\cdot 0 & 1\cdot 3 \\ 3\cdot 2 & 3\cdot 1 & 1\cdot 2 & 1\cdot 1 \end{bmatrix} = \begin{bmatrix} 0 & 3 & 0 & 6 \\ 2 & 1 & 4 & 2 \\ 0 & 9 & 0 & 3 \\ 6 & 3 & 2 & 1 \end{bmatrix}$

The dummy variables can be represented as

$$I_N = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, J_T = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

The reproduction of the dummy variables above amounts to

$I_N \otimes J_T = Z_\mu$

Our model can then be written as

$$y = \begin{bmatrix} Z_\mu & x \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \nu_{it}$$

The Kronecker product $(I_N \otimes J_T)$ is a block diagonal matrix and X is the matrix of nonconstant regressors

- An OLS estimator of this model yields the LSDV estimator

$$\begin{bmatrix} \hat{\alpha}_{LSDV} \\ \hat{\beta}_{LSDV} \end{bmatrix} = \begin{bmatrix} (I_N \otimes J_T)'(I_N \otimes J_T) & (I_N \otimes J_T)'x \\ x'(I_N \otimes J_T) & x'x \end{bmatrix}^{-1} \times \begin{bmatrix} (I_N \otimes J_T)'y \\ x'y \end{bmatrix}$$

$$= \begin{bmatrix} TI_N & T\bar{x} \\ T\bar{x}' & x'x \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{y} \\ x'y \end{bmatrix}$$

- The LSDV model can easily be estimated using over the full panel to yield LSDV estimators.
- This model is appealing
- But **for short panels** the problem is that it estimates too many (**incidental)** parameters that may not be of intrinsic value

  $K+1+(N-1)$

  K parameters for the original X-regressors;

  1 parameter for the intercept;

  N-1 parameters for cross-section fixed effects (omitted cross-section captured by the common intercept i.e the reference class-Tanzania.)

Problems with LSDV Model

1. There are too many incidental/nuisance parameters since μ_i grows as N increases. The usual proof of consistency for an estimator does not hold for LSDV model.

2. Inverting (K+1)+(N-1) matrix may be impossible if N is very large. Even when it is possible it can be inaccurate.

Eviews panel results of LSDV model

Dependent Variable: LN_RCONS?
Method: Pooled Least Squares
Date: 09/08/2011 Time: 21:11
Sample: 1990 2003
Included observations: 14
Cross-sections included: 12
Total pool (unbalanced) observations: 163

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LN_RGDP?	0.683031	0.050445	13.54021	0.0000
_BOTS--C	2.042464	0.502558	4.064134	0.0001
_BURKF--C	4.224775	0.718861	5.877042	0.0000
_BUR--C	4.730530	0.630225	7.506101	0.0000
_KEN--C	3.976501	0.681912	5.831401	0.0000
_MADAG--C	3.065602	0.508071	6.033802	0.0000
_MAURIT--C	3.215030	0.580446	5.538895	0.0000
_MOR--C	1.471438	0.295430	4.980671	0.0000
_NIG--C	2.076701	0.434629	4.778103	0.0000
_RWA--C	2.025546	0.325644	6.220130	0.0000
_SIERL--C	4.366440	0.721471	6.052135	0.0000
_RSA--C	3.813676	0.687518	5.547019	0.0000
_TAN--C	2.569156	0.444138	5.784585	0.0000

R-squared	0.998770	Mean dependent var	10.40442
Adjusted R-squared	0.998671	S.D. dependent var	2.939862
S.E. of regression	0.107162	Akaike info criterion	-1.552558
Sum squared resid	1.722550	Schwarz criterion	-1.305817
Log likelihood	139.5335	F-statistic	10147.81
Durbin-Watson stat	0.961912	Prob(F-statistic)	0.000000

Stata is not quite good in estimating the LSDV model. Eviews does a good job

We need a trick to deal with these problems

4.1.2 WITHIN/Q Estimator

- Using the "WITHIN" estimation we can still assume individual effects, although we no longer directly estimate them.
- We **demean** the data so as to "wipe out the incidental parameters (individual effects) and estimate β only.
- This means subtracting the mean for each cross-section from each observation.
- Demeaning the data will not change the estimates for β. (Think of the econometric exercise of "running a regression line through the origin".)

- In order to wipe out the individual effects, we define a Q matrix

$$Qy = Qx'\beta + Q\nu_{it}$$

Where $Q = I_N - P$

$$P = Z_\mu \left(Z'_\mu Z_\mu\right)^{-1} Z_\mu$$

- P is a centering matrix that averages across time for each individual cross-section
- Consequently, pre-multiplying this regression by Q obtains deviations from the means within each cross-section

$\widetilde{y} = Qy$ and $\widetilde{x} = Qx$

The OLS estimator is

$$\widetilde{\beta} = \left(x'Qx\right)^{-1} x'Qy$$

$$\text{var}\,\widetilde{\beta} = \sigma_v \left(x'Qx\right)^{-1}$$

Let's look at the same stuff using common parlance

$$y_{it} = \alpha + \mu_i + x'_{it}\beta + \nu_{it}$$

The mean model is

$$\bar{y}_{i\bullet} = \alpha + \mu_i + \bar{x}'_{i\bullet}\beta + \bar{\nu}_{i\bullet}$$

Demeaning the model

$$y_{it} - \bar{y}_{i\bullet} = \alpha + \mu_i + x'_{it}\beta + \nu_{it} - \left(\alpha + \mu_i \bar{x}'_{i\bullet}\beta + \bar{\nu}_{i\bullet}\right)$$

$$= \left(\alpha - \alpha\right) + \left(\mu_i - \mu_i\right) + \left(x'_{it} - x'_{i\bullet}\right)\beta + \left(\nu_{it} - \bar{\nu}_{i\bullet}\right)$$

$$y_{it} - \bar{y}_{i\bullet} = \left(x'_{it} - \bar{x}'_{i\bullet}\right)\beta + \left(\nu_{it} - \bar{\nu}_{i\bullet}\right)$$

Where

$$\bar{y}_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

$$\bar{x}_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$$

$$\bar{\nu}_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} \nu_{it}$$

Notice that we have wiped out the individual effect coefficients since

$$\left(\alpha - \alpha\right) = 0$$

$$\left(\mu_i - \mu_i\right) = 0$$

Using OLS yields the WITHIN estimator

$$\hat{\beta}_W = \left[\sum_{i=1}^{N}\sum_{i=1}^{T}\left(x_{it} - \bar{x}_{i\bullet}\right)\left(x_{it} - \bar{x}_{i\bullet}\right)'\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \bar{x}_{i\bullet}\right)\left(y_{it} - \bar{y}_{i\bullet}\right)$$

- There are no incidental parameters and the errors still satisfy the usual assumptions.

- We can therefore use OLS on the above equation to obtain consistent estimates.

- Averaging across all observations yields

$$\bar{y}_{\bullet\bullet} = \alpha + \beta \bar{x}_{\bullet\bullet} + \bar{v}_{\bullet\bullet}$$

- Individual effects can be solved (not estimated) with the assumption:

$\sum_{i=1}^{N} \mu_i = 0$ to avoid the dummy variable trap or perfect multicollinearity

and solving:

$$\tilde{\alpha} = \bar{y}_{\bullet\bullet} - \tilde{\beta}_1 \bar{y}_{\bullet\bullet} - \beta_2 \bar{x}_{\bullet\bullet}$$

$$\tilde{\mu}_{i\bullet} = \bar{y}_{i\bullet} - \tilde{\alpha} - \tilde{\beta}_1 \bar{x}_{i\bullet} - \beta_2$$

- In other words, we can use First Order Conditions to derive individual effects.

- Note that the total individual effect is the sum of the common constant **and** the constructed individual component.

		A	B	A-C
Country	**Period**	**Consumption**	**mean consumption**	**demeaned consumption**
Botswana	1998	7180.3	7797.021896	-616.7
Botswana	1999	7533.5	7797.021896	-263.6
Botswana	2000	7841.1	7797.021896	44.1
Botswana	2001	7919.2	7797.021896	122.2
Botswana	2002	8085.2	7797.021896	288.2
Botswana	2003	8222.9	7797.021896	425.9
Burkina Faso	1998	1283027.4	1406484.816	-123457.4
Burkina Faso	1999	1297642.2	1406484.816	-108842.6
Burkina Faso	2000	1306400.0	1406484.816	-100084.8
Burkina Faso	2001	1411715.4	1406484.816	5230.6
Burkina Faso	2002	1513999.2	1406484.816	107514.4
Burkina Faso	2003	1626124.7	1406484.816	219639.9
Burundi	1998	462066.7	477841.1	-15774.4
Burundi	1999	492055.5	477841.1	14214.4
Burundi	2000	465738.0	477841.1	-12103.1
Burundi	2001	469289.9	477841.1	-8551.2
Burundi	2002	491701.3	477841.1	13860.2
Burundi	2003	486195.2	477841.1	8354.1
Kenya	1998	596883.1	626869.0426	-29986.0
Kenya	1999	594332.1	626869.0426	-32537.0
Kenya	2000	609862.0	626869.0426	-17007.0
Kenya	2001	629103.7	626869.0426	2234.7
Kenya	2002	650968.4	626869.0426	24099.3
Kenya	2003	680065.0	626869.0426	53196.0
Madagascar	1998	21830.2	22222.41045	-392.2
Madagascar	1999	22441.8	22222.41045	219.4
Madagascar	2000	22483.0	22222.41045	260.6
Madagascar	2001	22443.8	22222.41045	221.4
Madagascar	2002	21150.2	22222.41045	-1072.2
Madagascar	2003	22985.4	22222.41045	763.0
Mauritius	1998	69552.9	75384.83324	-5831.9
Mauritius	1999	71594.9	75384.83324	-3789.9
Mauritius	2000	73939.3	75384.83324	-1445.5
Mauritius	2001	76048.7	75384.83324	663.9
Mauritius	2002	78570.9	75384.83324	3186.1
Mauritius	2003	82602.2	75384.83324	7217.4
Morocco	1998	240.3	248.4780447	-8.1
Morocco	1999	233.4	248.4780447	-15.0
Morocco	2000	243.0	248.4780447	-5.5
Morocco	2001	256.4	248.4780447	7.9
Morocco	2002	256.1	248.4780447	7.6
Morocco	2003	261.7	248.4780447	13.2
Nigeria	1998	3307.9	3028.129497	279.7
Nigeria	1999	2255.7	3028.129497	-772.4
Nigeria	2000	2446.5	3028.129497	-581.6
Nigeria	2001	3068.0	3028.129497	39.9
Nigeria	2002	3665.8	3028.129497	637.6
Nigeria	2003	3424.9	3028.129497	396.7
Rwanda	1998	588.0	668.6450459	-80.7
Rwanda	1999	595.6	668.6450459	-73.1
Rwanda	2000	641.9	668.6450459	-26.7
Rwanda	2001	676.1	668.6450459	7.5
Rwanda	2002	740.4	668.6450459	71.8
Rwanda	2003	769.9	668.6450459	101.2
Sierra Leone	1998	1180237.6	1254557.641	-74320.1
Sierra Leone	1999	1032168.8	1254557.641	-222388.8
Sierra Leone	2000	1142680.0	1254557.641	-111877.6
Sierra Leone	2001	1369830.5	1254557.641	115272.9
Sierra Leone	2002	1547871.3	1254557.641	293313.6
Sierra Leone	2003	1613277.6	1254557.641	358720.0
South Africa	1998	516925.9	566165.8325	-49239.9
South Africa	1999	531213.0	566165.8325	-34952.8
South Africa	2000	556652.0	566165.8325	-9513.8
South Africa	2001	579316.4	566165.8325	13150.5
South Africa	2002	598804.9	566165.8325	32639.1
South Africa	2003	614082.8	566165.8325	47917.0
Tanzania	1998	5610.4	6550.223129	-939.8
Tanzania	1999	6003.2	6550.223129	-547.0
Tanzania	2000	6069.6	6550.223129	-480.6
Tanzania	2001	6579.9	6550.223129	29.7
Tanzania	2002	7064.1	6550.223129	513.9
Tanzania	2003	7974.2	6550.223129	1423.9

Within estimation results from **Eviews**

Dependent Variable: LN_RCONS?
Method: Pooled Least Squares
Date: 09/08/08 Time: 21:16
Sample: 1990 2003
Included observations: 14
Cross-sections included: 12
Total pool (unbalanced) observations: 163

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.082438	0.540824	5.699521	0.0000
LN_RGDP?	0.683031	0.050445	13.54021	0.0000
Fixed Effects (Cross)				
_BOTS--C	-1.039974			
_BURKF--C	1.142337			
_BUR--C	1.648092			
_KEN--C	0.894063			
_MADAG--C	-0.016836			
_MAURIT--C	0.132592			
_MOR--C	-1.611000			
_NIG--C	-1.005736			
_RWA--C	-1.056892			
_SIERL--C	1.284002			
_RSA--C	0.731238			
_TAN--C	-0.513282			

Effects Specification

Cross-section fixed (dummy variables)

R-squared	0.998770	Mean dependent var	10.40442
Adjusted R-squared	0.998671	S.D. dependent var	2.939862
S.E. of regression	0.107162	Akaike info criterion	-1.552558
Sum squared resid	1.722550	Schwarz criterion	-1.305817
Log likelihood	139.5335	F-statistic	10147.81
Durbin-Watson stat	0.961912	Prob(F-statistic)	0.000000

- The fixed effects have not been computed but simply recovered
- This can be seen from the **lack of standard errors** as opposed to the LSDV results
- The interpretation of the country-specific fixed effects is as follows

- For those countries with positive values, it means that there are some unobservable factors which tend to enhance consumption
- For those countries with negative country-specific fixed effects, there are unobservable characteristics that hinder the consumption

Look at stata within results

```
xtreg ln_cons ln_rgdp, fe i(country)

Fixed-effects (within) regression               Number of obs      =       154
Group variable (i): country                     Number of groups   =        11

R-sq:  within  = 0.5644                         Obs per group: min =        14
       between = 0.9903                                        avg =      14.0
       overall = 0.9888                                        max =        14

                                                F(1,142)           =    183.96
corr(u_i, Xb)  = 0.9591                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_cons |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ln_rgdp |   .6576598   .0484889    13.56   0.000     .5618063    .7535133
       _cons |   3.255495   .5148909     6.32   0.000     2.237653    4.273337
-------------+----------------------------------------------------------------
     sigma_u |  1.0877784
     sigma_e |  .10204792
         rho |  .99127587   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(10, 142) =   127.58             Prob > F = 0.0000
```

Properties of the WITHIN Estimators

- The slope coefficients are consistent if N or T become large
- The fixed effects are only consistent if T is large
- The number of degrees of freedom must be adjusted. The degrees of freedom k=NT-N-K.
- **Please note that the usual OLS programs(software) not designed for panel data assume that the degrees of freedom k=NT-K, which is wrong!!!** Their standard errors, test statistics and p-values must be corrected as follows. Let's use the following notation $k_u = NT - K$ (unadjusted degrees of freedom) and $k_a = k_u - N$ (adjusted degrees of freedom)

$$se_a(\hat{\beta}) = \sqrt{\frac{k_u}{k_a} se_u(\hat{\beta})}$$

$$t_a(\hat{\beta}) = \sqrt{\frac{k_a}{k_u} t_u(\hat{\beta})}$$

Which is distributed t_{v_a} under the null hypothesis

Note that "a" denotes adjusted and "u" denotes unadjusted

- The parameter estimates from the LSDV are the same as from the WITHIN regression model.
- Please note that this is not a general result since incidental parameters do cause inconsistencies in may applied models.

Important disadvantage with the WITHIN method:

- Demeaning the data means that X-regressors which are dummy variables cannot be used.
- For example sex, race, religion, etc.
- Thus we would be able to say nothing about the relationship between the dependent variable and the time-invariant characteristics using this estimator.

5. RANDOM EFECTS MODEL (REM)

- The benefit of REM approach is that you concede variation across the cross-sections, but don't estimate "*N-1*" of these variations.

- However, in this approach you introduce a more complicated variance structure and OLS is no longer appropriate.

- This method is best suited to "random" draws from a large population ex. household surveys which claim to be "representative". (*N* is usually quite large.)

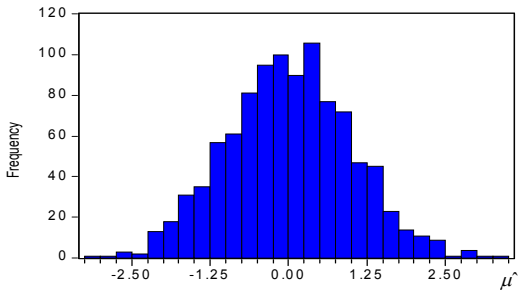
- The problems of too many parameters with LSDV model and "sweeping away" the time-invariant regressors "can be avoided" if μ_i are assumed random, i.e. drawn from a given distribution.

$$\mu_i \sim IID(0, \sigma_\mu^2)$$

$$v_{it} \sim IID(0, \sigma_v^2)$$

and μ_i are independent of v_{it}.

- In other words we are assuming that the individual effects have an empirical distribution function.



Which has certain characteristics.

$\alpha = average \qquad \mu = \frac{1}{N}\sum_{i=1}^{N}\mu_i$

$\sigma_\mu^2 =$ Variance of μ

- We can use these definitions to write the panel data model in form of REM

$$y_{it} = \alpha + x'_{it}\beta + (\mu_i - \alpha) + \nu_{it}$$

The new error term is $u_{it} = (\mu_i - \alpha) + \nu_{it}$

We can then rewrite the REM model as;

$$y_{it} = \alpha + x'_{it}\beta + u_{it}$$

This is almost like the pooled model, except for the following;

- The constant term can be interpreted as the average individual effects
- The error term has a special complicated form

- We can estimate the REM model using OLS to obtain estimates of α and β.

- These estimates will only be consistent if the following conditions hold;
  - $E(u_{it}) = E(\mu_i - \alpha) + E(\nu_{it}) = 0$
  - $Cov(u_{it}, x_{it}) = Cv(\mu_i, x_{it}) + Cov(\nu_{it}, x_{it}) = 0$, i.e. no correlation between individual effects and regressors

5.1 Efficiency in the REM

For REM to be efficient, two conditions must fulfilled;

- Homoscedasticity : $Var(u_{it}) = \sigma_\mu^2 + \sigma_v^2$ for all *i* and *t*. Here we assume that μ_i and u_{it} are independent
- Serial independence in the error term , $u_{it} = (\mu_i - \alpha) + v_{it}$

  $Cov(u_{it}, u_{js}) = \sigma_\mu^2 + \sigma_v^2$ if $i = j$ and $s = t$ (Same cross-section, same year)

  $Cov(u_{it}, u_{js}) = 0$ if $i \neq j$ and $s = t$ If individuals are independent)

  $Cov(u_{it}, u_{js}) = \sigma_\mu^2 \neq 0$ if $i = j$ and $s \neq t$ (Same cross-section, different year)

The last condition violates the serial independence assumption.

OLS is thus inefficient in a REM and thus yields incorrect standard errors and tests.

5.2 Feasible GLS (FGLS) Estimator

$y_{it} = \alpha + x_{it}'\beta + u_{it}$

- The FGLS estimator for the REM can be implemented by OLS regression of the transformed equation as follows

1. Define $\hat{\theta} = 1 - \frac{\sigma_v}{\sigma_1}$ where $\sigma_1^2 = T\sigma_\mu^2 + \sigma_v^2$
2. Calculate "pseudo within differences"

$y_{it}^* = y_{it} - \hat{\theta}\bar{y}_{i\bullet}$, $x_{it}^* = x_{it} - \hat{\theta}\bar{x}_{i\bullet}$

3. Perform an OLS regression

$$y_{it}^{*} = \alpha^{*} + \beta x_{it}^{*} + u_{it}^{*}$$

Where $\alpha^{*} = (1-\hat{\theta})\alpha$ and $u_{it}^{*} = (1-\hat{\theta})\alpha_i + (\varepsilon_{it} - \hat{\theta}\varepsilon_i)$

4. The REM estimate of β is given by;

$$\hat{\beta}_{re} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}^{*} - \bar{x}_{i\bullet}^{*})(y_{it}^{*} - \bar{y}_{i\bullet}^{*})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}^{*} - \bar{x}_{i\bullet}^{*})^{2}}$$

The estimator of the intercept can be shown to equal

$\mu_{re} = \bar{y} - \hat{\beta}_{re}\bar{x}$ (Greene, 2010)

5.2.1 The Crucial Problem: We do not know θ

- The unfortunate thing is that we do not know σ_{μ}^{2} and σ_{v}^{2}
- If the errors u_{it}, v_{it} and μ_i were known, we could estimate the variances easily as follows;

1. $\hat{\sigma}_{1}^{2} = \frac{T}{N}\sum_{i=1}^{N}\bar{u}_{i\bullet}$

2. $\hat{\sigma}_{v}^{2} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - \bar{u}_{i\bullet})^{2}$

   $= \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}(v_{it} - \bar{v}_{i\bullet})^{2}$

3. $\hat{\sigma}_{\mu}^{2} = \frac{1}{(N-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mu_i - \bar{\mu})^{2}$

- Since u_{it}, v_{it} and μ_i are unknown, there are a number of suggestions on how they can be estimated.

- These methods use various residuals instead of unknown parameters.

Possible Residuals

1. $\hat{u}_{ols}$=REM residuals from the **pooled** regression $y_{it} = \alpha + x'_{it}\beta + u_{it}$, number of observations is NT

2. $\hat{u}_b$=REM residuals from the **between** regression $\bar{y}_{i\bullet} = \alpha + \bar{x}'_{i\bullet}\beta + \bar{u}_{i\bullet}$, number of observations is N.

3. $\hat{v}_w$=FE residuals from the **Within** regression $\tilde{y}_{it} = \tilde{x}'_{it}\beta + \tilde{v}_{it}$, number of observations is NT.

4. $\hat{u}_w$=REM residuals from the **within** regression. This can be computed as $\hat{v}_w + \left(\hat{\mu}_w - \bar{\hat{\mu}}_w\right)$.

5. $\hat{u}_{re}$=REM residuals from the regression $y^*_{it} = \alpha^* + \beta x^*_{it} + u^*_{it}$, number of observations is NT.

1. Wallace and Hussein (1969)

Use $\hat{u}_{ols}$ (unbiased and consistent but not efficient) instead of u in 4 and 5

2. Swamy and Arora (1972)

Use $\hat{u}_b$ in 4 and $\hat{v}_w$ in 5. This is the approach used in Eviews econometric software when you estimate a REM. It is also the default in stata when you use **re** option.

3. Amemiya(1971)

Use $\hat{u}_w$ in 4 and $\hat{v}_w$ in 5

4. **Nerlove (1971)**

Use $\hat{\mu}_w$ in 6 and $\hat{v}_w$ in 5

5. **Wansbeek-Kapteyn (1989)** for incomplete panels

Some Comments

- There is no much difference between the models when the REM specification is correct
- Only Nerlove (1971) guarantees that $\sigma_\mu^2 > 0$. Many users of the other methods set $\theta = 1$ (fixed effects model) if a negative value of σ_μ^2 is found.
- There are no general rules as to which method to use. The most common is the Swammy-Arora(also used in Eviews)
- The REM estimates are more efficient when REM specification is correct. They are inconsistent when the model is incorrect
- It is important to test which model is correct

```
xtreg ln_cons ln_rgdp, re i(country)

Random-effects GLS regression                   Number of obs      =       154
Group variable (i): country                     Number of groups   =        11

R-sq:  within  = 0.5644                         Obs per group: min =        14
       between = 0.9903                                        avg =      14.0
       overall = 0.9888                                        max =        14

Random effects u_i ~ Gaussian                   Wald chi2(1)       =    879.56
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ln_cons |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ln_rgdp |   .8916337   .0300645    29.66   0.000     .8327083     .950559
       _cons |   .7713074   .3363147     2.29   0.022     .1121428    1.430472
-------------+----------------------------------------------------------------
     sigma_u |  .31717419
     sigma_e |  .10204792
         rho |  .90619338   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

5.3 BETWEEN Estimator

- The between estimator uses just the cross-sectional variation.
- For instance for a model $y_{it} = \alpha_i + x'_{it}\beta + \nu_{it}$
- We could average all the years to yield $\bar{y}_i = \alpha_i + \bar{x}'_i\beta + \bar{\nu}_i$
- This can be rewritten as a between model

$\bar{y}_i = \alpha + \bar{x}'_i\beta + (\alpha_i - \alpha + \nu_i)$

Where $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, $\bar{\varepsilon}_i = T^{-1}\sum_{t=1}^{T} \varepsilon_{it}$ and $\bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}$

- The between estimator is the OLS estimator for regression of $\bar{y}_i$ on time averaged regressors
- The concern is the difference between different individuals (i.e. "between estimator") and is the analogue of cross-section regression which is a special case T=1
- For instance for our consumption example, the data for consumption will be as follows

Country	Mean real consumption
Botswana	6926.8
Burkina Faso	1166573.8
Burundi	545623.9
Kenya	547946.8
Madagascar	20678.3
Mauritius	64759.6
Morocco	234.0
Nigeria	2878.3
Rwanda	618.8
Sierra Leone	1376062.0
South Africa	500589.7
Tanzania	5386.2

Do the same for gdp

- The between estimator is consistent if the regressors $\bar{x}_i$ are independent of the composite error $(\alpha_i - \alpha + \nu_i)$

- This will be case for the constant-coefficients model and the REM model.
- For the fixed effects model, the between estimator is inconsistent as α_i is assumed to be correlated with x_i and hence $\bar{x}_i$

```
Between regression (regression on group means)  Number of obs      =       154
Group variable (i): country                     Number of groups   =        11

R-sq:  within  = 0.5644                         Obs per group: min =        14
       between = 0.9903                                        avg =      14.0
       overall = 0.9888                                        max =        14

                                                F(1,9)             =    920.77
sd(u_i + avg(e_i.))=  .3183446                  Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_cons |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ln_rgdp |   .9996306   .0329431    30.34   0.000     .9251081    1.074153
       _cons |   -.375336   .3627005    -1.03   0.328    -1.195821    .4451495
------------------------------------------------------------------------------
```

5.3.1 Comments

- The REM estimator $\hat{\beta}_{re}$ of the slope parameters converge to the within estimator as $T \to \infty$ and as $\theta \to 1$
- We can show that the GLS estimator is a weighted average of the within and between estimator

$$\hat{\beta}_{re} = w_1 \beta_{within} + w_1 \hat{\beta}_{be}$$

5.4 FEM vs. REM MODEL

- The FEM vs REM is an issue that has generated a hot debate in panel econometrics literature

(1) Traditional criterion: ***Is*** μ_i ***viewed as a random variable or as parameter to be estimated?***

μ_i *is* ***"random effect"-*** when the individual effects are randomly distributed across the cross-sections

μ_i is **"fixed effect"-** when it is treated as a parameter to estimated for each cross-section observation i.

(2) *Modern panel data econometrics-The key issue here is whether* μ_i *is correlated with the regressors or not.*

μ_i *is random effect-****When there is zero correlation between the observed explanatory variables and the*** μ_i $i.e. Cov(X_{it}, \mu_i) = 0$

μ_i **is fixed effect-**when there is correlation between the observed explanatory variables and μ_i .

In other words, we allow for arbitrary correlation between the unobserved effects μ_i and the observed explanatory variables

5.4. 1 Individual Specific Variables (Time-invariant regressors)

- Many times when we conduct research we have exogenous variables that vary between individuals but which do not vary over time within a given individual (e.g.gender, race, language, nationality etc.).

- These are called time-invariant regressors

Which is the best model to deal with such exogenous variables?

To understand this let's denote individual specific variables as s_i

In a FEM, we will write it as follows;

$$y_{it} = \alpha + \mu_i + \lambda s_i + x'_{it}\beta + \nu_{it}$$

Let's look at the two FEM approachs: LSDV and WITHIN estimation

a) **LSDV**: The LSDV will not be able to estimate it because of perfect multicollinearity between the individual-specific effects, μ_i, and the individual specific variables s_i. The reason is because in both cases dummy variables are used

b) **WITHIN:** Under this approach the term $(\alpha + \mu_i + \lambda s_i)$ does not vary over time and will thus be removed by within transformation (demeaning process). $y_{it} - \bar{y}_{i\bullet} = \beta_1(x_{it} - \bar{x}_{i\bullet}) + (\nu_{it} - \bar{\nu}_{i\bullet})$

This means that the parameters of the individual specific variables cannot be estimated using the FEM. This means that we cannot distinguish between observed and unobserved heterogeneity, a feature that may be important for policy.

In REM we will write the model as follows

$$y_{it} = \alpha + \lambda s_i + x'_{it}\beta + u_{it}$$

In this case λ can easily be estimated, although not when using the Nerlove method. It is however, important to note that for the REM to be appropriate, the observed heterogeneity, s_i, must be independent of the unobserved heterogeneity, μ_i.

The REM thus offers an added advantage by allowing us to estimate parameters for time-invariant regressors which may be of policy relevance.

Note however, that in the context of the REM, we cannot interpret the coefficients for the unobserved heterogeneity!!.

Look at eviews output

Dependent Variable: LN_RCONS?
Method: Pooled EGLS (Cross-section random effects)
Date: 09/08/08 Time: 22:56
Sample: 1990 2003
Included observations: 14
Cross-sections included: 12
Total pool (unbalanced) observations: 163
Swamy and Arora estimator of component variances

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.019244	0.371701	2.742110	0.0068
LN_RGDP?	0.879092	0.032618	26.95085	0.0000
Random Effects (Cross)				

_BOTS--C	-0.922566
_BURKF--C	0.411866
_BUR--C	1.256634
_KEN--C	0.307806
_MADAG--C	0.074454
_MAURIT--C	-0.057190
_MOR--C	-0.687420
_NIG--C	-0.625203
_RWA--C	-0.253274
_SIERL--C	0.542767
_RSA--C	0.124030
_TAN--C	-0.171903

Effects Specification

	S.D.	Rho
Cross-section random	0.419148	0.9386
Idiosyncratic random	0.107162	0.0614

Weighted Statistics

R-squared	0.812596	Mean dependent var	0.721526
Adjusted R-squared	0.811432	S.D. dependent var	0.266552
S.E. of regression	0.115749	Sum squared resid	2.157038
F-statistic	698.1075	Durbin-Watson stat	0.954621
Prob(F-statistic)	0.000000		

Unweighted Statistics

R-squared	0.964211	Mean dependent var	10.40442
Sum squared resid	50.10865	Durbin-Watson stat	0.041094

We cannot interpret the coefficients of the random effects coefficients because they are randomly distributed across the cross-sections

Question:

Suppose you have a one-way error components model of the form $y_{it} = \alpha + \mu_i + x'_{it}\beta + \nu_{it}$

Prove that the sum of μ_i is equal to zero.

Solution

The question is equivalent to $\sum_{i=1}^{N} \mu_i = 0$

Use the fact that $\mu_i = \alpha_i - \alpha$

$$\sum_{i=1}^{N} \mu_i = \sum_{i=1}^{N} (\alpha_i - \alpha) = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha = N\alpha - N\alpha = 0$$

References

Baltagi, B.H. (2009), *Econometric Analysis of Panel Data*, 3rd Edition, John Wiley Chapter 1

Cameron,A.C. and Trivedi, P.K. (2006), *Microeconometrics: Methods and Applications*, Cambridge University Press. Chapter Chapter 21

Cheng Hsiao, *Analysis of Panel Data*, Cambridge University Press, 2005, Chapter 11

Greene W H. *Econometric Analysis*, Second Edition, Macmillan, 2010. Chapter 13

Wooldridge, J. M. (2002), *Econometric Analysis of Cross-Section and Panel Data*, MIT Press: Cambridge, Massachusets. Chapter 10