Linear Panel Data Models with Computer Applications

By
S.Madheswaran, ISEC, Bangalore

Types of Data Sets

- Cross Section: Data on one or more variables relating to many (similar) cross sectional units at a particular point of time (to study cross sectional behavior)
- Time Series: Data on one or more variables relating to one unit over time (to study the dynamic behavior)
- Panel Data: Data relating to many cross section units over time (to study both dynamic and cross sectional behavior)

Cross Section Data (2003)

Public Sector Banks	Int. Margin	Deposits	Borrow	Fixed Assets	Employees
STATE BANK OF BIKANER & JAIPUR	5513769	132336276	3105353	1015911	5758
STATE BANK OF HYDERABAD	7477606	205989353	4164215	1153140	13396
STATE BANK OF MYSORE	3866052	90131158	3354081	382934	10443
STATE BANK OF PATIALA	7890422	178696752	4339854	1193933	11667
STATE BANK OF SAURASHTRA	3195149	90509712	4438924	431935	7404
STATE BANK OF TRAVANCORE	5228475	159262799	484102	722680	11327
STATE BANK OF INDORE	3671810	92168078	3002808	467338	6546
STATE BANK OF INDIA	99775570	2961232824	93036194	23885483	209797
BANK OF BARODA	21033639	663663655	6253308	6973198	40249
ALLAHABAD BANK	9097760	254633836	436924	3576223	19761
BANK OF INDIA	20362046	644535959	40269286	7367012	43198
BANK OF MAHARASHTRA	6763100	221757503	3148227	1554148	14052
CANARA BANK	22329501	720948225	938223	6596163	45084
DENA BANK	5681091	164912588	2282114	2961308	10573
INDIAN BANK	8203878	270159286	4492114	4238604	22215
INDIAN OVERSEAS BANK	12214634	366985910	3559653	2925135	24476
CENTRAL BANK OF INDIA	18974253	511651192	1469905	7523148	39678
UNION BANK OF INDIA	14976807	447486187	4420730	7332877	25707
PUNJAB NATIONAL BANK	31237136	758134973	6621645	8847021	58895
UNITED BANK OF INDIA	7196825	210312921	577180	1873773	17242
SYNDICATE BANK	12097269	306605432	787709	3431814	26472
ANDHRA BANK	7529687	210618474	9906275	1681017	13000
CORPORATION BANK	7921439	217245740	8033408	2329477	10246
PUNJAB & SIND BANK	3862636	132236249	247586	684167	8860
ORIENTAL BANK OF COMMERCE	12047600	298090879	7660179	1452806	13440
VIJAYA BANK	6433896	170198109	3208178	1593881	11168

Time Series Data

YEAR	Bank	Int. Margin	Deposits	Borrow	Fixed Asst.	Employees
2000	SBI	692835	19682107	927807	247761	233433
2001	SBI	824779	24282838	1072203	259330	214845
2002	SBI	908125	27056014	932394	241523	209462
2003	SBI	997756	29612328	930362	238855	208998
2004	SBI	1118632	31861867	1343133	264512	207039
2005	SBI	1394463	36704753	1918431	269769	205515

Panel Data

YEAR	Banks	Int. Margin	Deposits	Borrow	Fixed Asst.	Employees
2000	ABN AMRO BANK	24188	342293	311350	8065	854
2001	ABN AMRO BANK	33137	460950	276693	7737	910
2002	ABN AMRO BANK	34742	486529	138909	8566	959
2003	ABN AMRO BANK	36625	502230	289050	8932	986
2004	ABN AMRO BANK	44912.5	585644.3	292493.39	8660.66	1207
2005	ABN AMRO BANK	57358.46	702598.27	496617.83	8405.12	1908
2000	ABU DHABI COMMERCIAL BANK	1409	59048	1000	798	71
2001	ABU DHABI COMMERCIAL BANK	1853	169716	4700	805	71
2002	ABU DHABI COMMERCIAL BANK	1698	166251	2441	872	82
2003	ABU DHABI COMMERCIAL BANK	1656	176865	1000	812	78
2004	ABU DHABI COMMERCIAL BANK	1630	181598	0	787	78
2005	ABU DHABI COMMERCIAL BANK	1190	166255	11843	759	78
2000	ALLAHABAD BANK	56422	1764210	3458	32916	22125
2001	ALLAHABAD BANK	68086	2010602	7018	36018	21009
2002	ALLAHABAD BANK	73047	2266594	6057	36813	19860
2003	ALLAHABAD BANK	90977	2546338	4369	35762	19515
2004	ALLAHABAD BANK	108574.87	3147660.5	16899.31	35304.96	19284
2005	ALLAHABAD BANK	136403.86	4076207.43	12948.98	73199.95	19309
2000	AMERICAN EXPRESS BANK	10654	141834	88412	7601	874
2001	AMERICAN EXPRESS BANK	9546	137429	178852	8342	972
2002	AMERICAN EXPRESS BANK	9920	104822	195119	8241	1105
2003	AMERICAN EXPRESS BANK	10864	238793	30000	8264	1367
2004	AMERICAN EXPRESS BANK	12736	278852	13201	7891	1479
2005	AMERICAN EXPRESS BANK	11853	226442	15375	6900	1479
2000	ANDHRA BANK	41564	1441795	14640	7083	14603
2001	ANDHRA BANK	50048	1829152	17793	9446	12798
2002	ANDHRA BANK	57535	1849077	21904	10854	12812
2003	ANDHRA BANK	75297	2106185	99063	16810	12991
2004	ANDHRA BANK	91058	2294052	84300	17993	13095
2005	ANDHRA BANK	106904	2755071	98324	18729	13107

Panel Data Sets

- A data set with both a cross section and a time dimension
- Also called as Longitudinal or Pooled cross section and time series data
- Examples:
  - Annual Survey of Industries Data (ASI)
  - CMIE DATA BASE (BROWESS)
  - Banking Statistics (RBI)
  - State Finance Data (RBI)

Panel (or) Longitudinal Data

- Pooling cross section wise time series data
- It has both cross sectional as well as time variations
- Balanced Panel: Same number of observations on each unit, so that the total number of observations is n.T (we consider the case where n>T)

Unbalanced Panel: Some observations missing for a few time period **(no special techniques are required)**

Advantages of Panel Data

- Many data points/observations

$\Rightarrow$More information and more DF

- Reduce Multi-collinearity

- Enable to study the complex dynamic behavior (Time and individual variations in behavior; unobservable in cross sections or aggregate time series)

Advantages…

- Avoids aggregation problems

- It provides a means of resolving / reducing the magnitude of a key econometric problem, namely the effects of missing or unobserved variables (omitted variable bias). It helps us to control for individual heterogeneity

Omitted Variable Bias

- If omitted variables are time invariant, we can get reliable estimates!
- Consider the true model $y_{it} = \alpha + \beta x_{it} + \gamma Z_i + \mathbf{u}_{it}$
- Unfortunately, we cannot measure z_i.
- It is "*lurking*" or "*latent*." By considering the changes

$$y_{it}^* = y_{it} - y_{i,t-1} = (\alpha + \beta x_{it} + \gamma z_i + u_{it}) - (\alpha + \beta x_{it-1} + \gamma z_i + u_{it-1}) = \beta(x_{it} - x_{it-1}) + (u_{it} - u_{it-1})$$
$$= \beta x_{it}^* + u_{it}^*$$

we do not need to worry about the bias that ordinarily arises from the latent variable, z_i.

- By introducing the subject-specific variable α_i, we can capture the impacts of many latent variables.

Panel data models: A Simple example

Let we have two variables: Y and X

Example: Y is Output and X Input (capital stock)

Objective: Analyze the impact of X on Y

Let we have data on Y and X for 15 major banks for 7 years (n.T=105)

How do we model it?

Two ways of modeling

Year B1 B2……… . B15 Y X Y X……….Y X	Banks T1 T2………….T7 Y X Y X……….Y X
1 2 . . 7 **A**	1 2 . . 15 **B**

A: 15 time series data

B: 7 Cross Section Data

Model A and Model B

- Assuming linear relation, estimable equation:

A: $Y_t = \alpha + \beta X_t + u_t$ (TS model for each bank)

B: $Y_i = \alpha + \beta X_i + u_i$ (Cross sect.model for each year)

- For A, we need to run 15 equations
- For B, we need to run 7 equations
- Let in both cases, all αs and βs are the same, then we can pool the data to have pooled (panel) data, and Pooled OLS is used to estimate single eqn.:

$Y_{it} = \alpha + \beta X_{it} + u_{it}$

Let αs vary

A	B
varying α means that they are certain unobserved heterogeneity or characteristics of banks influence. So, we need to incorporate those effects in our model by bank specific attributes $Y_{it} = \alpha + \beta X_{it} + \lambda_i + u_{it}$ One way (cross sectional) effect model	varying α means that years are not the same: some years are normal and some are boom or depression. So, we need to incorporate those effects in our model by time specific variable $Y_{it} = \alpha + \beta X_{it} + \mu_t + u_{it}$ One way (time) effect model

How to model λ_i and/or μ_t

- λ_i: Dummies for each cross sectional units
- μ_t : dummies for each time period
- They may treated as (i) explanatory variables or (ii) residuals
- In regression, explanatory variables are non-stochastic or fixed while residuals are random or stochastic
- If fixed, then fixed effects model estimation and if random, then random effects method.

One way Effect Model

- If we consider only one effect it is called the one way effect model
- Assume initially that we have one way effect (i.e., cross sectional heterogeneity)

Two way Effects Model

- If we incorporate both cross sectional heterogeneity and time variations, the model is two ways model

To decide whether we can pool the data, testing for Stability of Regression

- The Model using matrix notation:

$$y = x\,\beta + \varepsilon \text{ (cross section/time series)}$$

- Before pooling the data, we have to test the hypothesis known as the stability of the regression across firms/time.

- Chow Test

Null H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_N$ (β' s are same)

Alternate H_1 : $\beta_1 \neq \beta_2 \neq \beta_3 \ldots\ldots\ldots \neq \beta_N$

Chow Test: Steps

- Run the regression: $y = x\beta + u$ for each sample firm separately (OLS);

- Then run a pooled regression (OLS)

$$F = \frac{(e'e - e'_1 e_1 - e'_2 e_2 - \ldots\ldots\ldots e'_N e_N) / (N-1)K'}{(e_1' e_1 + e_2' e_2 + \ldots\ldots\ldots e_N' e_N) / N(T-K')}$$

This is distributed as $F[(N-1)K, N(T-K)]$

- Note: One can also test the stability of regression across time {d.f = [(T-1)K, N(T-K)]}.

Chow Test for testing significance of the Group Effects

- Similar Test as above, but between the pooled regression and pooled regression including only firm dummies or time dummies

- $$F(n-1, nT-n-K) = \frac{(R^2_{LSDV} - R^2_{Pooled})/(n-1)}{(1-R^2_{LSDV})/(nT-n-K)}$$

- Alternatively:

- $$F(n-1, nT-n-K) = \frac{(RSS_{pooles} - RSS_{LSDV})/(n-1)}{(1-R^2_{LSDV})/(nT-n-K)}$$

INDIVIDUAL REGRESSIONS

- When both intercept and slope vary across firms but not over time:

  One can use separate equation for each firm . If error term in each equation satisfies usual assumptions and if there is no contemporaneous correlation between the errors of two equations, then these equations are unrelated. By applying OLS to each equation, we can obtain BLUE estimators.

SUR Model

Sometimes, common events that occur in any economy often affect the different cross sectional errors in a similar way so that they are contemporaneously correlated. For instance, $Cov(u_{1t}, u_{2t}) \neq 0 = \sigma_{12}$ and so on.

So we have to apply the seemingly unrelated regression (SUR). It is the term coined by Zellner (1962) because they appear unrelated except for the correlations among the residuals.

Step 1: Each equation is estimated separately by OLS and calculate residuals.

Step 2: Use the residuals estimated from Step 1 to calculate variances and co-variances

Step 3: Use the estimates from Step 2 to obtain GLS estimates of parameters.

Fixed Effects Model

- Suppose we specify the model as:

$Y_{it} = \alpha + \beta X_{it} + \delta_1 D_1 + \delta_2 D_2 + \; + \delta_3 D_3 + \ldots.$
$\ldots\ldots\ldots\ldots\ldots\ldots + \delta_n D_n + u_{it}$

- Dummy variable Trap Problem
- Remedy: Drop α or one of the dummies.
- LIMDEP will include n dummies, dropping α.
- Modified Equation: $Y_{it} = \alpha_i [i] + \beta X_{it} + \varepsilon_{it}$
- Note: Here, no time-specific effects.
- Simple LSDV Model: OLS can be used

Within and Between Groups Estimators

3 ways of formulating panel reg. models:

Pooled : $y_{it} = \alpha + \beta' x_{it} + \varepsilon_{it}$

within : $y_{it} - \bar{y}_i = \beta' [x_{it} - \bar{x}_i] + \varepsilon_{it} - \bar{\varepsilon}_i$

Between: $\bar{y}_i = \alpha + \beta' \bar{x}_i + \bar{\varepsilon}_i$

For (1), the moments would be about the overall means.

For (2), the moment matrices are within-groups sums of squares and cross products.

For (3), the moments are the between-groups sums of squares and cross products

Within and Between Groups Estimators

- Three possible Least Squares Estimators of β corresponding to the decomposition

- They are: b^T , b^W and b^B

- It can be shown that the OLS estimator is a matrix weighted average of the within and between groups estimators

Between Estimation

$$\overline{y_{i\cdot}} = X_{i\cdot}\beta + \text{error}$$

where the ith term $\overline{y_{i\cdot}}$ is

$$\overline{y_{i\cdot}} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

$$\widehat{\beta_B} = (X'P_D X)^{-1} X'P_D y$$ where $$P_D = D(D'D)^{-1}D',$$

Within Estimation

$$y_{it} - \overline{y_{i\cdot}} = (X_{it} - \overline{X_{i\cdot}})\beta + \text{error}$$

$$\widehat{\beta_W} = (X'M_D X)^{-1} X'M_D y$$ where $$M_D = I_{nT} - D(D'D)^{-1}D',$$

Pooled OLS

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}(X'M_D y + X'P_D y) \\ &= (X'X)^{-1}X'M_D X\hat{\beta}_W + (X'X)^{-1}X'P_D X\hat{\beta}_B\end{aligned}$$

Exercise 1

- Data on X and Y for Three Firms (1, 2 and 3) during 1990 to 2004 (15 years) are given.
- **Run the individual regression (Y on X)**
- **Run the pooled regression**
- **Run the dummy variable regression**
- **Run within group estimation**
- **Plot (Scatter) the data and add trend lines**
- **Compare the results**

Exercise 2

- Dep: Ln q
- Independent: Ln x
- Run Pooled Regression
- Run Within Estimation
- Run Fixed Effects Model

Fixed Vs. Random Effects Models

- The fixed effects approach takes α_i to be a group specific constant term while the random effects approach specifies that α_i is a group specific disturbance

- Random Effects model assumes that unobserved individual specific effect is uncorrelated with X variable and that randomly distributed across firms which are from a large population.

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it} + u_{i}\ \_\ \text{a random}$$

- Also called the error component model

- OLS can not be used?

OLS Vs. Feasible GLS

- In OLS, the estimated Coefficient Vector is:

  β (cap) = $(x' x)^{-1} x' y$

- GLS Estimates are:

  β (cap) = $(x' \Omega^{-1} x)^{-1} x' \Omega^{-1} y$

Random Effects Model Estimation Procedure

$$y_{it} = X_{it}\beta + \epsilon_{it}$$

$$\epsilon_{it} = \alpha_i + \eta_{it}$$

$$E[\eta] = 0 \qquad E[\eta\eta'] = \sigma_\eta^2 I_{nT}$$

$$E[\alpha_i \alpha_j] = 0, \text{ for } i \neq j \qquad E[\alpha_i \alpha_i] = \sigma_\alpha^2$$

$$E[\alpha_i \eta_{jt}] = 0 \qquad E[\alpha_i] = 0$$

$$E[\epsilon_i \epsilon_i'] = \sigma_\eta^2 I_T + \sigma_\alpha^2 ii' = \begin{bmatrix} \sigma_\eta^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\eta^2 + \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\eta^2 + \sigma_\alpha^2 \end{bmatrix}$$

$$\Omega = I_n \otimes \Sigma = E[\epsilon\epsilon'] = \begin{bmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \Sigma \end{bmatrix}$$

Whether to choose fixed or random

- Hausman's (1978) Specification Test for the RE model (for orthogonolity of REs and regressors)
- It rests on the idea that under the null H. of no correlation [$E(u_{it}|X_{it}) = 0$], both OLS, LSDV and GLS are consistent but OLS is inefficient. Under alternative OLS/LSDV is consistent and GLS is not.
- Under the null two estimates should not differ and a test can be based on the difference.
- It is a χ^2 test based on the Wald criterion

Wald Statistics

$$W = (b^{GLS} - b^{lsdv})' \, \Sigma^{-1} \, (b^{GLS} - b^{lsdv}) \sim \chi^2(k)$$

where $\Sigma = \mathrm{var}\,(b^{GLS} - b^{lsdv})$ and

k - number of regressors in X

Hausman Test

- If W < the table χ^2 value for appropriate d.f and level of significance, the null of individual effects are uncorrelated with other regressors can not be rejected (accepted).
- In this case, the Random Effects model is relevant (and not the Fixed Effects Model).
- Simple Rule: Larger W favors for Fixed effects model and lower value for random effects model

Testing For Random Effects (LM Test) : Breusch-Pagan Test

- It is based on the OLS residuals
- Null Hypothesis: $\sigma^2_v = 0$
- Alternate Hypothesis: $\sigma^2_v \neq 0$
- Let $e'e$ be the RSS from OLS
- $LM_c = [nT/(2(T-1)] = [(T^2 e'e/e'e)-1]^2 \sim \chi^2(1)$
- If $LM_c > LM_T$, reject the null hypothesis and the model supports the random effects model
- Note: When T is large, this test is not valid.

Fixed Time Effects and Group Effects

- $Y_{it} = \alpha_i + \beta X_{it} + \mu_t + u_{it}$

  where, μ = T-1 time dummies

- As usual Chow Test can be used to see whether time effect is significant.

- If yes, Two way model is relevant

Two Way Fixed Effects

- Includes Individual as well as time dummies.

  $y_{it} = \alpha_i + \mu_t + x_{it}'\beta + \varepsilon_{it}$

  - Normalization needed as the individual and time dummies both sum to one.
  - Reformulate model:

    Add T-1 dummies
  - Complication: Unbalanced panels are difficult

THANK YOU