Panel Regression in Stata

An introduction to type of models and tests

Gunajit Kalita
Rio Tinto India

STATA Users Group Meeting
1st August, 2013, Mumbai

Content

- Understand Panel structure and basic econometrics behind
- Application of different Panel regression models and post estimation tests in STATA

What are Panel Data?

Panel data are a type of *longitudinal data*, or data collected at different points in time. Three main types of *longitudinal data*:

- **Time series data**: Many observations (large t) on as few as one unit (small N). Examples: stock price trends, aggregate national statistics
- **Pooled cross sections:** Two or more independent samples of many units (large N) drawn from the same population at different time periods:
    - General Social Surveys
    - India's Decennial Census
- **Panel data:** Two or more observations (small t) on many units (large N)
    - Panel surveys of households and individuals (NSS EUS, CES)
    - Data on organizations and firms at different time points (ASI, NSS)
    - Aggregated country/regional data over time (WDI,WEO,BOP)
- The literature on econometrics of panel regression and options available in STATA is vast, this presentation will only introduce the fundamentals of this topic today

Advantage of Panel Data

Heterogeneity	• It relate to individuals, firms, states, countries etc., over time, presence of heterogeneity in these units is natural • Such heterogeneity can be explicitly taken into account by allowing individual specific variables
Degree of freedom	• It combines time series of cross section observations, thus • Gives more informative data, more variability, less collinearity among variables, more degree of freedom and more efficiency • By studying repeated cross section of observation, it is better suited to study dynamics of change
Unobservable	• Panel data can better detect and measures effects that simply can not be observed in pure cross section or time series data. • For example, the effect of minimum wage laws on employment and earnings can be better studied if we include successive waves of minimum wage increase in the federal and/or state minimum wages
Behavioural Models	• Panel data enables us to study more complicated behavioural models • For example, phenomenon such as economies of scale and technological change can be better handled by panel data • It can also minimise the bias that might result if we aggregate individuals or firms into broad aggregates

Data requirement

- Basic panel methods require at least two "waves" of measurement

  Consider services share of GDP in a country and its economic development (GDP per capita) in the last three decades

- One way to construct your panel is to create a single record for each combination of unit (country, firm, individual) and time period

- Data include:
  - ***A time-invariant*** unique identifier for each unit (country, firm, individual)
  - ***A time-varying*** outcome (Services share in GDP, GDP, Inflation)
  - ***An indicator of time*** (Year, Quarter, Month, day)

- Variation for dependent variable and regressors:
  Overall: Over time and individuals
  Between: Between individuals
  Within: Within individuals (over time)

Country Name	Year	Services, etc., value added (% of GDP)	GDP per capita, PPP (constant 2005 international $)	GDP per capita (constant 2000 US$)	Urban population (% of total)	Population, total	Age dependency ratio (% of working-age population)	Trade (% of GDP)
India	1992	44.9	1238	316.8	25.9	910064576	70.38	18.60
India	1993	45.2	1272	325.5	26.2	928226051	69.66	19.90
India	1994	44.7	1330	340.5	26.4	946373316	68.90	20.30
India	1995	45.7	1404	359.4	26.6	964486155	68.11	23.10
India	1996	45.6	1482	379.4	26.8	982553253	67.29	22.20
India	1997	47.1	1515	387.7	27	1000558144	66.45	22.90
India	1998	47.9	1580	404.5	27.3	1018471141	65.57	24.00
India	1999	49.7	1684	426.9	27.5	1036258683	64.68	25.30
India	2000	50.5	1722	436.6	27.7	1053898107	63.77	27.40
India	2001	51.5	1778	451.9	27.9	1071374264	62.84	26.40
India	2002	52.7	1818	461.5	28.1	1088694080	61.90	30.00
India	2003	52.8	1932	492.4	28.3	1105885689	60.96	30.90
India	2004	53.0	2052	525	28.5	1122991192	60.04	36.90
India	2005	53.0	2209	565.3	28.7	1140042863	59.14	41.30
India	2006	52.9	2378	608.7	29	1157038539	58.26	45.30
India	2007	52.7	2573	658.8	29.3	1173971629	57.42	44.90
India	2008	54.2	2635	681.5	29.5	1190863679	56.60	52.70
India	2009	55.3	2813	733.1	29.8	1207740408	55.82	44.90
India	2010	54.7	3039	786.7	30.1	1224615000	55.06	46.30
Indonesia	1992	41.7	2270	669.1	32.6	190512441	64.62	52.80
Indonesia	1993	42.4	2396	706.5	33.6	193525648	63.34	50.50
Indonesia	1994	42.1	2538	748.3	34.6	196488446	62.06	51.90
Indonesia	1995	41.1	2711	799.3	35.6	199400339	60.77	54.00
Indonesia	1996	39.9	2877	848.2	36.9	202257039	59.45	52.3
Indonesia	1997	39.6	2971	876.0	38.2	205063468	58.13	56.0
Indonesia	1998	36.7	2547	750.8	39.4	207839287	56.85	96.2
Indonesia	1999	37.0	2533	746.8	40.7	210610776	55.69	62.9
Indonesia	2000	38.5	2623	773.3	42	213395411	54.67	71.4
Indonesia	2001	38.3	2683	791.1	43.2	216203499	53.81	69.8
Indonesia	2002	40.1	2768	816.0	44.4	219026365	53.08	59.1
Indonesia	2003	41.1	2863	844.2	45.7	221839235	52.44	53.6
Indonesia	2004	41.0	2970	875.7	46.9	224606531	51.85	59.8
Indonesia	2005	40.3	3102	914.6	48.1	227303175	51.25	64.0
Indonesia	2006	40.1	3236	953.9	49.2	229918547	50.65	56.7
Indonesia	2007	39.5	3403	1003.4	50.3	232461746	50.05	54.8
Indonesia	2008	37.5	3570	1052.4	51.5	234951154	49.46	58.6

Panel data models

Pooled Model

- The pooled model specifies constant coefficients, the usual assumptions for cross-sectional analysis. It is most restrictive panel model

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

- The default standard errors erroneously assume errors are independent over *i* for given *t.*

Individual-specific effects model

- We assume that there is unobserved heterogeneity across individuals captured by α_i

  Example: unobserved ability of an individual that affects wages

- The main question is whether the individual-specific effects α_i are correlated with the regressors.
- If they are correlated, we have the ***fixed effects (FE) model***. If they are not correlated we have the ***random effects (RE) model***

Individual-specific effects model

Fixed effects model (FE)

- It allows individual-specific effects α_i to be correlated with the regressors x . We include α_i as intercepts. Each individual has a different intercept term and the same slope parameters

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

- We can recover the individual specific effects after estimation as:

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

In other words, the individual-specific effects are the leftover variation in the dependant variable that cannot be explained by the regressors

Random effects model (RE)

- It assumes that individual-specific effects are distributed independently of the regressors, we include α_i in the error term. Each individual has the same slope parameters and a composite error term $\varepsilon_{it} = \alpha_i + e_{it}$

$$y_{it} = x_{it}^{'}\beta + (\alpha_i + e_{it})$$

- Here $\text{var}(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_e^2$ and $\text{cov}(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2$, so $\rho_\varepsilon = cor(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_e^2)$
- Rho is the ***interclass correlation*** of the error. Rho is the fraction of the variance in the error due to the individual-specific effects. It approaches 1 if the individual effects dominate the idiosyncratic error

Choosing between fixed and random effects

Breusch-Pagan Lagrange Multiplier (LM) test

- This is a test for the random effects model based on the OLS residual. The LM test helps to decide between a random effects regression and a simple OLS regression
- The null hypothesis is that variances across entities is zero. Test whether σ_u^2 or equivalently $cor(u_{it}, u_{is})$ is significantly different from zero.
- If the LM test is not significant, it implied no significant difference across units(i.e. no panel effect), thus can run simple OLS regression

Hausman test

- The null hypothesis is that the preferred model is random effects vs. the alternative fixed effects. It tests whether the unique errors (α_i) are correlated with the regressors, the null hypothesis is they are not correlated.
- The random effects estimator is more efficient so we need to use it if the Hausman test supports it. The Hausman test statistic can be calculated only for the time-varying regressors
- The Hausman test statistic is:

$$H = \left(\hat{\beta}_{RE} - \hat{\beta}_{FE}\right)' \left(V\left(\hat{\beta}_{RE}\right) - V\left(\hat{\beta}_{FE}\right)\right) \left(\hat{\beta}_{RE} - \hat{\beta}_{FE}\right)$$

Example: Cross country panel

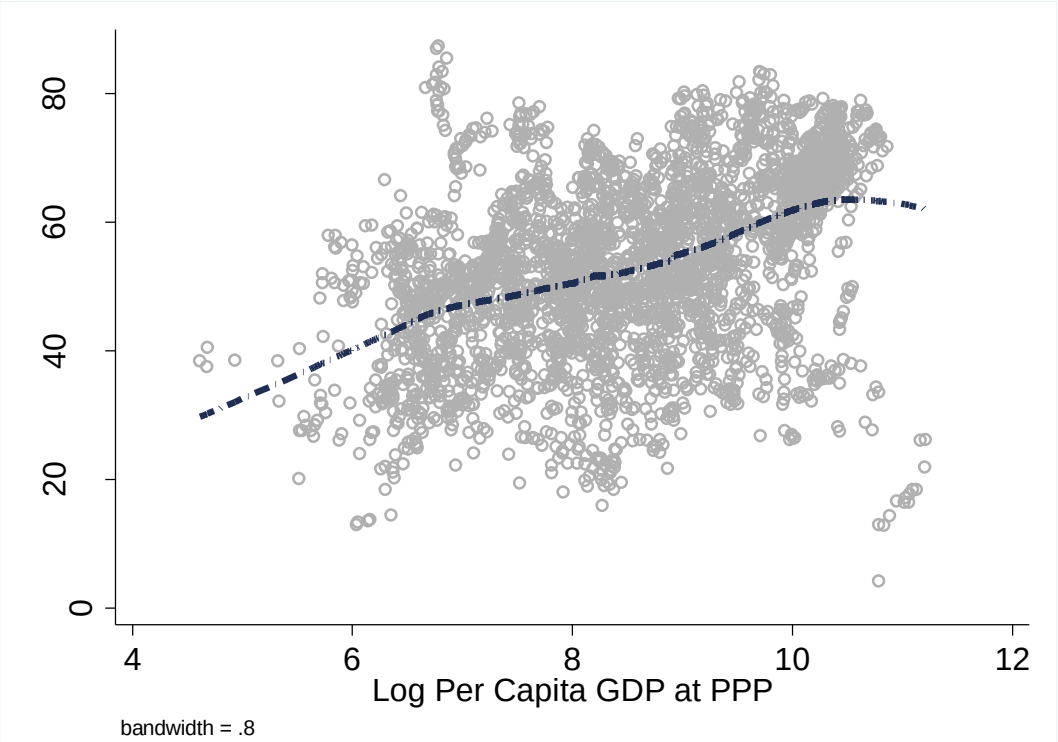
Two Waves of Services Growth (NBER WP:14968)

"The positive association between the service sector share of output and per capita income is one of the best-known regularities in all of growth and development economics. Yet there is less than complete agreement on the nature of that association. Here we identify two waves of service sector growth…"

- They identify two waves of service sector growth, a first wave in countries with relatively low levels of per capita GDP and a second wave in countries with higher per capita incomes
- There is evidence of the second wave occurring at lower income levels after 1990
- **Does that mean India's experience is not an aberration?**

Command: lowess ser_sh lngdpc_pp

Lowess Plot of the Relationship between Log Per Capita Income and Services/GDP (1980-2010), 116 countries



$$\frac{Serv_{it}}{GDP_{it}} = \text{Constant} + \sum_i \theta_i D_i + \alpha_1 Y_{it} + \alpha_2 Y_{it}^2 + \alpha_3 Y_{it}^3 + \alpha_4 Y_{it}^4 + \varepsilon_{it}$$

Panel-Fixed effect (FE) model

STATA Commands:

- To convert country name from string to individual code
  encode country, gen(con_cod)
- Declare the Panel variables
  xtset con_code year
- Run country fixed effect model
  xtreg ser_sh lngdpc_pp lngdp_pp2 lngdp_pp3 lngdp_pp4 lngdp_90s lngdp_20s,fe

```
Fixed-effects (within) regression               Number of obs      =      3397
Group variable: con_cod                         Number of groups   =       113

R-sq:  within  = 0.1984                         Obs per group: min =        10
       between = 0.2191                                        avg =      30.1
       overall = 0.2085                                        max =        31

                                                F(6,3278)          =    135.22
corr(u_i, Xb)  = 0.1150                         Prob > F           =    0.0000
```

ser_sh	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lngdpc_pp	332.9264	74.79444	4.45	0.000	186.2779	479.575
lngdp_pp2	-60.60611	14.42567	-4.20	0.000	-88.89036	-32.32187
lngdp_pp3	4.906946	1.213405	4.04	0.000	2.527837	7.286054
lngdp_pp4	-.1477659	.0376061	-3.93	0.000	-.2214997	-.0740322
lngdp_90s	.3742022	.0312394	11.98	0.000	.3129514	.435453
lngdp_20s	.6419146	.0370546	17.32	0.000	.5692621	.7145671
_cons	-642.7124	142.5919	-4.51	0.000	-922.2906	-363.1343
sigma_u	10.953122					
sigma_e	5.8722998					
rho	.77673786	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(112, 3278) =   101.25          Prob > F = 0.0000
```

Panel-Random effect (RE) model

STATA Commands:

- Run random effect model
  xtreg ser_sh lngdpc_pp lngdp_pp2 lngdp_pp3 lngdp_pp4 lngdp_90s lngdp_20s,re

- Testing for cross-sectional dependence or contemporaneous correlation
  xtcsd, pesaran abs

```
Random-effects GLS regression                   Number of obs      =      3397
Group variable: con_cod                         Number of groups   =       113

R-sq:  within  = 0.1983                         Obs per group: min =        10
       between = 0.2220                                        avg =      30.1
       overall = 0.2130                                        max =        31

                                                Wald chi2(6)       =    841.07
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

ser_sh	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lngdpc_pp	352.3767	73.52802	4.79	0.000	208.2644	496.489
lngdp_pp2	-64.61057	14.17162	-4.56	0.000	-92.38643	-36.83472
lngdp_pp3	5.26195	1.191796	4.42	0.000	2.926072	7.597828
lngdp_pp4	-.1590866	.0369467	-4.31	0.000	-.2315008	-.0866725
lngdp_90s	.3669355	.0308193	11.91	0.000	.3065308	.4273402
lngdp_20s	.6244614	.0347734	17.96	0.000	.5563067	.692616
_cons	-677.8364	140.3619	-4.83	0.000	-952.9406	-402.7321
sigma_u	10.817956					
sigma_e	5.8722998					
rho	.7724016	(fraction of variance due to u_i)				

Ho: Residual are not correlated

```
Pesaran's test of cross sectional independence =    16.947, Pr = 0.0000

Average absolute value of the off-diagonal elements =  0.439
```

OLS or RE or Fe

STATA Commands:

- **Breusch-Pagan Lagrange Multiplier (LM) test: OLS vs RE**

  quietly *xtreg ser_sh lngdpc_pp lngdp_pp2 lngdp_pp3 lngdp_pp4 lngdp_90s lngdp_20s,re*
  xttest0

- **Hausman test: RE vs FE**

  quietly *xtreg ser_sh lngdpc_pp lngdp_pp2 lngdp_pp3 lngdp_pp4 lngdp_90s lngdp_20s,fe*
  estimate store fe

  quietly *xtreg ser_sh lngdpc_pp lngdp_pp2 lngdp_pp3 lngdp_pp4 lngdp_90s lngdp_20s,re*
  estimate store re
  hausman fe re

Breusch and Pagan Lagrangian multiplier test for random effects

ser_sh[con_cod,t] = Xb + u[con_cod] + e[con_cod,t]

Estimated results:

	Var	sd = sqrt(Var)
ser_sh	191.0374	13.82163
e	34.48391	5.8723
u	117.0282	10.81796

Test: Var(u) = 0

chibar2(01) = 29076.72

Prob > chibar2 = 0.0000

	Coefficients (b) fe	(B) re	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
lngdpc_pp	332.9264	352.3767	-19.45025	13.70544
lngdp_pp2	-60.60611	-64.61057	4.00446	2.695435
lngdp_pp3	4.906946	5.26195	-.3550045	.2279756
lngdp_pp4	-.1477659	-.1590866	.0113207	.0070114
lngdp_90s	.3742022	.3669355	.0072667	.0051062
lngdp_20s	.6419146	.6244614	.0174533	.0128005

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 4.58

Prob>chi2 = 0.3337

Prediction

STATA Commands:

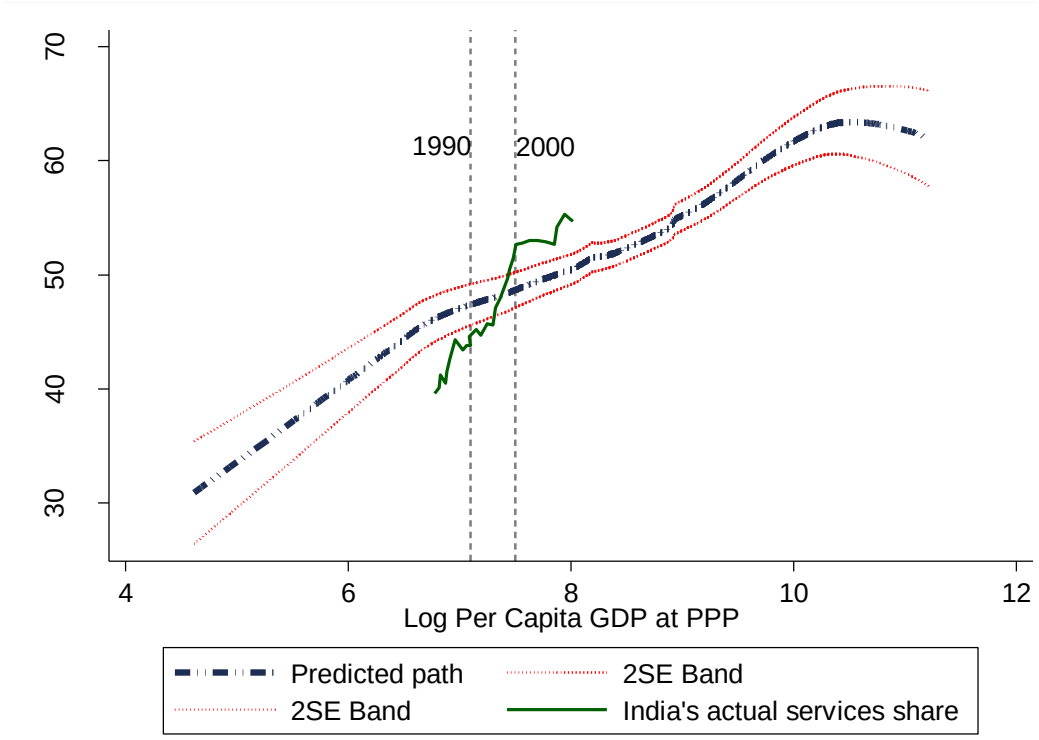
- Prediction fitted value including individual-specific effects
  predict yhat, xbu

- Prediction standard error of the fitted values
  predict se, stdp

- Prediction standard error band
  *gen up_se=yhat+2*se*
  *gen low_se=yhat-2*se*

- Lowess Curve

  twoway (lowess yhat lngdpc_pp)(lowess up_se lngdpc_pp) (lowess low_se lngdpc_pp)(line ser_sh lngdpc_pp if (con_cod==50))



To produce robust standard error estimates for linear panel models

Command	Option	SE estimates are robust to disturbances being	Notes
`reg, xtreg`	`robust`	heteroscedastic	
`reg, xtreg`	`cluster()`	heteroscedastic and autocorrelated	
`xtregar`		autocorrelated with AR(1)[1]	
`newey`		heteroscedastic and autocorrelated of type MA(q)[2]	
`xtgls`	`panels(), corr()`	heteroscedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1)	$N < T$ required for feasibility; tends to produce optimistic SE estimates
`xtpcse`	`correlation()`	heteroscedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1)	large-scale panel regressions with `xtpcse` take a lot of time
`xtscc`		heteroscedastic, autocorrelated with MA(q), and cross-sectionally dependent	

[1] AR(1) refers to first-order autoregression

[2] MA(q) denotes autocorrelation of the moving average type with lag length q.

References

- Panel data analysis, Princeton University, http://dss.princeton.edu/training/
- Econometric Academy by Ani Katchova, https://sites.google.com/site/econometricsacademy/econometrics-models
- Introduction to Regression Models for Panel Data Analysis, Indiana University by Prof. Patricia A. McManus, http://www.indiana.edu/~wim/docs/10_7_2011_slides.pdf
- Econometric analysis using Panel Data by Ranjit Kumar Paul, http://www.iasri.res.in/sscnars/socialsci/12-Panel%20data.pdf
- Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence by Daniel Hoechle, http://fmwww.bc.edu/repec/bocode/x/xtscc_paper.pdf
- Two Waves of Services Growth by Poonam Gupta and Barry Eichengreen, NBER Working Paper no. 14968, http://www.nber.org/papers/w14968.pdf

Thank You

Gunajit Kalita

Gunajit.kalita@riotinto.com

gunajitkalita1984@gmail.com

My Blog: http://macroscan.wordpress.com/