Labor Economics with STATA

Liyousew G. Borga

February 3, 2016

Panel Data Methods

Outline

Panel Data

- Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities are observed across time.
- These entities could be states, companies, individuals, countries, etc.

Panel Data

- Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities are observed across time.
- These entities could be states, companies, individuals, countries, etc.

Advantages of Panel Data

- Allow us to identify causal effects under weaker assumptions
  - you can control for variables you cannot observe or measure (e.g. cultural factors)
  - you can control for variables that change over time but not across entities (e.g. policies)
- Allow us to study individual trajectories
  - variables at different levels of analysis (i.e. students, schools, districts, states)
  - transitions into and out of states (e.g. poverty; employment)

Panel Data

- Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities are observed across time.
- These entities could be states, companies, individuals, countries, etc.

Advantages of Panel Data

- Allow us to identify causal effects under weaker assumptions
  - you can control for variables you cannot observe or measure (e.g. cultural factors)
  - you can control for variables that change over time but not across entities (e.g. policies)
- Allow us to study individual trajectories
  - variables at different levels of analysis (i.e. students, schools, districts, states)
  - transitions into and out of states (e.g. poverty; employment)

Drawbacks of Panel Data

- data collection issues (i.e. sampling design, coverage), non-response (e.g. in micro panels), or cross-country dependency in macro panels (i.e. correlation between countries)

Panel data sets come in two forms:

- Balanced panel: each cross-sectional unit is observed for the same time periods
- Unbalanced panel: cross-sectional units are observed for different amounts of time

Some terminology:

- A short panel has a large number of individuals but few time observations on each
- A long panel has a long run of time observations on each individual, permitting separate time-series analysis for each
- Attrition is the process of drop-out of individuals from the panel, leading to an unbalanced and possibly non-compact panel

Endogeneity Problem

- Suppose you are interested in understanding the linear relationship between x and y using the following linear model:

$$y_i = \alpha + \beta x_i + u_i$$

- The most critical assumption of a linear regression is the exogeneity assumption (i.e. the error term and the regressor must be statistically independent): $E(u_i|x_i) = 0$
- $E(u_i) = 0$: The (unconditional) mean of the error term is 0
- $Cov(x_i, u_i) = 0$: The error term does not correlate with X
- In many non-experimental social science research settings the exogeneity assumption will be violated ($E(u_i|x_i) \neq 0$)
- The X variation that is used to identify the causal effect is endogenous

Example

- Cross-section earnings regression

$$y_i = \mathbf{z_i}\alpha + \mathbf{x_i}\beta + \varepsilon_i$$

where:

y_i = log wage;

z_i = observable time-invariant factors (education, etc.);

x_i = observable time-varying factors (e.g. job tenure);

ε_i= random error (e.g. "luck")

Possible misspecifications, causing bias:

- Omitted dynamics (lagged variables not observed)
- Reverse causation (e.g. pay and tenure jointly determined)
- Omitted unobservables (e.g. "ability")

Example: Identification of unobservables

$$y_{it} = \mathbf{z_i}\alpha + \mathbf{x_{it}}\beta + u_i + \varepsilon_{it}$$

where u_i= unobservable "ability" (assumed not to change over time)

- Pooled data regression of y on $\mathbf{z}$ and $\mathbf{x}$ will suffer from omitted variable bias
- Ability, u, is likely to be positively related to education, $\mathbf{z}$: Pooled OLS will result in an upward bias in the estimated of returns to education

How do we identify the effect of u_i if we can't observe it?

Example: Identification of unobservables

$$y_{it} = \mathbf{z_i}\alpha + \mathbf{x_{it}}\beta + u_i + \varepsilon_{it}$$

where u_i= unobservable "ability" (assumed not to change over time)

- Pooled data regression of y on $\mathbf{z}$ and $\mathbf{x}$ will suffer from omitted variable bias
- Ability, u, is likely to be positively related to education, $\mathbf{z}$: Pooled OLS will result in an upward bias in the estimated of returns to education

How do we identify the effect of u_i if we can't observe it?

- depends on the assumptions about the correlation structure of the compound residual: $v_{it} = u_i + \varepsilon_{it}$
- If individuals (i and j) have been sampled at random:

$$cov(u_i, u_j) = 0$$

$$cov([\varepsilon_{i1} \ldots \varepsilon_{iT}], [\varepsilon_{j1}\varepsilon_{jT}]) = 0$$

- But there may be some correlation over time for any individual for two different periods $s \neq t$:

$$cov(v_{is}, v_{it}) \neq 0$$

Example: Identification of unobservables

$$y_{it} = \mathbf{z_i}\alpha + \mathbf{x_{it}}\beta + u_i + \varepsilon_{it}$$

- Add and subtract an arbitrary combination of the **z**-variables ($\mathbf{z_i}\gamma$):

$$y_{it} = \mathbf{z_i}\alpha + \mathbf{z_i}\gamma + \mathbf{x_{it}}\beta + u_i - \mathbf{z_i}\gamma + \varepsilon_{it} \quad (1)$$

$$y_{it} = \mathbf{z_i}\alpha^* + \mathbf{x_{it}}\beta + u_i^* + \varepsilon_{it} \quad (2)$$

where: $\alpha^* = (\alpha + \gamma)$ and $u_i^* = (u_i - \mathbf{z_i}\gamma)$

- But (1) and (2) have exactly the same form, so we can't tell whether we are estimating α or a completely arbitrary value $\alpha^* = (\alpha + \gamma)$

Rubin's Causal Model

- According to the counterfactual approach to causality (Rubin's model) an individual causal effect is defined as

$$\Delta_i = Y^T_{i,t_0} - Y^C_{i,t_0} \qquad \text{(T: Treatment; C: Control)}$$

However, this is not estimable (fundamental problem of causal inference)

- Cross-sectional data: We compare different persons i and j (between estimation)

$$\hat{\Delta}_i = Y^T_{i,t_0} - Y^C_{j,t_0}$$

Assumption: unit homogeneity (no unobserved heterogeneity)

- Panel data: We compare the same person over time t_0 and t_1 (within estimation)

$$\hat{\Delta}_i = Y^T_{i,t_1} - Y^C_{i,t_0}$$

Assumption: temporal homogeneity (no period effects, no maturation)

- Panel data: Within estimation with control group

$$\hat{\Delta}_i = (Y^T_{i,t_0} - Y^C_{ij,t_0}) - (Y^T_{j,t_0} - Y^C_{j,t_0})$$

Assumption: parallel trends

Pooled OLS

$$y_{it} = \mathbf{z_i}\alpha + \mathbf{x_{it}}\beta + u_i + \varepsilon_{it}$$

- A pooled regression of y on $\mathbf{z}$ and $\mathbf{x}$ using all the data together would assume that there is no correlation across individuals, nor across time periods for any individual
- This would ignore the individual effect u; if u_i is correlated with $\mathbf{z_i}$ and $\mathbf{x_{it}}$, pooled regression is biased
- If u_i is uncorrelated with $\mathbf{z_i}$ and $\mathbf{x_{it}}$, pooled regression gives unbiased but inefficient results, with incorrect standard errors, t-ratios

Random Effects Methods

$$y_{it} = \mathbf{x_{it}}\beta + u_i + \varepsilon_{it}$$

- RE model assumes that the u_i are i.i.d. random-effects: No time-constant unobserved heterogeneity; No time-varying unobserved heterogeneity
- The random effects approach accounts for the serial correlation in the composite error $v_{it} = u_i + \varepsilon_{it}$
- We can therefore apply GLS methods that account for the particular error structure in $v_{it} = u_i + \varepsilon_{it}$

Fixed Effects Methods

$$y_{it} = \mathbf{x_{it}}\beta + u_i + \varepsilon_{it} \tag{1}$$

- In many applications the whole point of using panel data is to allow for arbitrary correlations of u_i with $\mathbf{x_{it}}$
- Fixed effects explicitly deals with the fact that u_i may be correlated with $\mathbf{x_{it}}$
- The fixed effects transformation: First, take the time means for each individual:

$$\bar{y}_i = \bar{\mathbf{x}}_{\mathbf{i}}\beta + u_i + \bar{\varepsilon}_i \tag{2}$$

- subtract (2) from (1), to get

$$y_{it} - \bar{y}_i = (\mathbf{x_{it}} - \bar{\mathbf{x}}_{\mathbf{i}})\beta + (u_i - u_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$\tilde{y}_i = \tilde{\mathbf{x}}_{\mathbf{i}}\beta + \tilde{\varepsilon}_i$$

- u_i drops out of the equation because it is time invariant

Fixed Effects Methods

- In FE models there are 3 ways to eliminate u_i
    - Within-transformation (FE transformation)
    - Least squares dummy variables
    - First differencing
- FE estimation amounts to controlling for every single time-invariant characteristic, observable or non-observable.
- This means, any time-invariant characteristic becomes irrelevant in determining β
- However, there may be time-variant unobservable characteristics (captured by $\tilde{\varepsilon}_i$; hence, consistency of β still requires x not to be correlated with $\tilde{\varepsilon}_i$

Fixed Vs Random Effects Methods

- RE more efficient than FE because RE also models person specific time-invariant effects (uses more information)
- RE not useful for causal effects of time-variant regressors
- RE or FE? To decide between FE and RE, estimate both models, run a Hausman test

	Fixed Effect Model	Random Effect Model
Functional form	$y_{it} = (\alpha + u_i) + \mathbf{x_{it}}\beta + \varepsilon_{it}$	$y_{it} = \alpha + \mathbf{x_{it}}\beta + (u_i + \varepsilon_{it})$
Assumption	-	Individual effects not correlated with regressors
Intercepts	Varying across group and/or time	Constant
Error variances	Constant	Randomly distributed across group and/or time
Slopes	Constant	Constant
Estimation	LSDV, within effect estimation	GLS, FGLS (EGLS)
Hypothesis test	F test	Breusch-Pagan LM test

Reading in Panel Data

Long versus Wide data sets

- long form: each observation is an individual-time (i, t) pair
- wide form: each observation is data on i for all time periods
- wide form: each observation is data on t for all individuals

The vast majority of Stata commands work best when the data is in long format

- `xt` commands require data in long form; use reshape long command to convert from wide to long form

```
reshape wide stub, i(id) j(time) //to convert formats from long to wide
reshape long stub, i(id) j(time) //to convert formats from wide to long
```

Load and Summarize Panel Data

Example Data

- We will use an artificial data on "Marital Wage Premium" (source: J. Bruderl, 2015)

```
use "Wage Premium.dta", clear
xtset id time
    panel variable:  id (strongly balanced)
    time variable:  time, 1 to 6
    delta:  1 unit
```

The note "(strongly balanced)" refers to the fact that all countries have data for all years

```
list id time wage marr, separator(6) // Listing the data
xtdes //panel description of the dataset
xtsum  //Panel summary statistics: within and between variation
```

Stata lists three different types of statistics: overall, between, and within

- Overall statistics are ordinary statistics that are based on 24 observations
- "Between" statistics are calculated on the basis of summary statistics of 4 individuals regardless of time period,
- "Within" is summary statistics of 6 time periods regardless of individuals

Plotting the data

```
twoway  (scatter wage time, ylabel(0(1000)5000, grid angle(0))    ///
         ymtick(500(1000)4500, grid) c(L))              ///
         (scatter wage time if marr==1, c(L)),           ///
         legend(label(1 "before marriage") label(2 "after marriage"))
```

Do married men earn more? Is There a Marriage-Premium for Men?

- Treatment between $t = 3$, and $t = 4$ (only for the two high-wage earners)
- There is a causal effect: a marriage-premium
- And we have a problem with self-selection: Only high-wage men marry
- The assumption of unit homogeneity does not hold

Panel data estimation: POLS

```
regress  wage marr if time==4 //cross-sectional regression
regress  wage marr       //POLS estimation with incorrect default S.E.
regress  wage marr, vce(cluster id)   //POLS with correct panel-robust S.E.
```

	Cross-section	Pooled OLS	POLS Robust
marriage	2500	1833.3***	1833.3
	(707.107)	(472.314)	(655.918)
Constant	1500	2166.7***	2166.7*
	(500.000)	(236.157)	(610.039)
Observations	4	24	24
Adjusted R^2	0.793	0.379	0.379

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- Cross-sectional estimation suffers from endogeneity
- The default standard errors in POLS erroneously assume errors are independent over i for given t
- Cluster-robust standard errors are much larger than the default. Always use cluster-robust s.e if use POLS

Within estimators

```
regress D.(wage marr), noconstant // First differences
```

- FD identifies the causal effect under weaker assumptions
- only time-varying unobserved heterogeneity must not be present
- with $t > 2$ FD-estimation is obviously inefficient

```
xtreg wage marr, fe       // Fixed-effects
```

- FE uses only within variation (of the treated only); the causal effect is identified by the deviations from the person-specific means

```
regress wage marr ibn.id, noconstant    // LSDV
```

- Practical only when N is small

Random effects

```
xtreg wage marr, re theta // Random effects
```

- RE estimator is a weighted average of the estimates produced by the between and within estimators

FE vs RE

- Prefer RE as can estimate all parameters and more efficient.
- But RE is inconsistent if fixed effects present
- Use Hausman test to choose between FE and RE
- This tests difference between FE and RE estimates is statistically significantly different from zero

Diagnostics

Testing for time-fixed effects

- To see if time fixed effects are needed when running a FE model use the command `testparm`

```
xtreg wage3 marr i.time, fe
testparm i.time
```

- If we fail to reject the null that the coefficients for all years are jointly equal to zero, then no time fixed-effects are needed

Testing for random effects

- The LM test helps you decide between a random effects regression and a simple OLS regression
- use the command `xttset0` right after running the random effects model

Testing for heteroskedasticity

- A test for heteroskedasticiy is available for the fixed- effects model using the command `xttest3`

```
ssc install xtest3
xttest3
```

Testing for serial correlation

- A Lagrange-Multiplier test for serial correlation is available using the command `xtserial`

```
ssc install xtserial
xtserial y x1
```

Testing for cross-sectional dependence

- cross-sectional dependence is a problem in macro panels with long time series
- less of an issue in micro panels (few years and large number of cases)

```
xtreg y x, fe
xttest2
```