**PANEL DATA ANALYSIS**

**MODELS FOR PANEL DATA**

Pane 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.

 **Example of panel data:**

 ** **  **   **

 **1 2010 3 3 6 2 1**

 **1 2011 2 7 2 5 8**

 **1 2012 5 6 8 1 3**

 **2 2010 2 4 2 5 6**

 **2 2011 6 2 3 1 4**

 **2 2012 3 1 2 8 5**

 **3 2010 1 2 4 3 6**

 **3 2011 3 4 7 8 2**

 **3 2012 2 3 6 9 3**

**Terminology and notations:**

Individual or cross section unit: country, region, state, firm, consumer, individual, couple of individuals or countries (gravity models) etc.

Double index : i (for cross-section unit) and t (for time)

for  and ****

**A micro-panel data set** is a panel for which the time dimension T is largely less important than the individual dimension N (example: Company X Panel Study of Income Dynamics, Compny B with 15,000 individuals observed since 1998):



**A macro-panel data set** is a panel for which the time dimension T is similar to the individual dimension N (example: a panel of 100 countries with quarterly data since the world war II):

 

A panel is said to be balanced if we have the same time periods, for each cross section observation. For an unbalanced panel, the time dimension, denoted *Ti* , is specific to each individual.

**Remark:** While the mechanics of the unbalanced case are similar to the balanced case, a careful treatment of the unbalanced case requires a formal description of why the panel may be unbalanced, and the sample selection issues can be somewhat subtle. => issues of sample selection and attrition.

**Advantages of using Panel Data**

Panel data sets for economic research possess several major advantages over conventional cross-sectional or time-series data

1. Panel data usually give the researcher a large number of data points (*N \* T*), increasing the degrees of freedom and reducing the collinearity among explanatory variables. Hence improving the efficiency of econometric estimates => but this kind of argument may not necessarily imply with more data points yields more information (heterogeneity bias).

2. More importantly, longitudinal data allow a researcher to analyze a number of important economic questions that cannot be addressed using cross-sectional or time-series data sets.

3. Panel data allows to control for omitted (unobserved) variables like cultural factors or difference in business practices across companies. Panel data provides a means of resolving the magnitude of econometric problems that often arises in empirical studies, namely the often heard assertion that the real reason one finds (or does not find) certain effects is the presence of omitted (unobserved) variables that are correlated with explanatory variables.

4. Panel data also help to control for unobservable variables that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc.)

5. Panel data involve two dimensions: a cross-sectional dimension N, and a time-series dimension T. We would expect that the computation of panel data estimators would be more complicated than the analysis of cross-section data alone (where T = 1) or time series data alone (where N = 1). However, in certain cases the availability of panel data can actually simplify the computation and inference.

6. With panel data you can include variables at different levels of analysis (i.e. students, schools, districts, states) suitable for multilevel or hierarchical modeling.

**Techniques Used to Analyze Panel Data**

1. Fixed effects (FE)

2. Random effects (RE)

**1. Fixed effects**

Fixed-effects (FE) explore the relationship between predictor and outcome variables within an entity (country, person, company, etc.).

Each entity has its own individual characteristics that may or may not influence the predictor variables (for example being a male or female could influence the opinion toward certain issue or the political system of a particular country could have some effect on trade or GDP or the business practices of a company may influence its stock price).

When using FE we assume that something within the individual may impact or bias the predictor or outcome variables and we need to control for this. This is the rationale behind the assumption of the correlation between entity’s error term and predictor variables. FE removes the effect of those time-invariant characteristics from the predictor variables so we can assess the predictors’ net effect.

Another important assumption of the FE model is that those time-invariant characteristics are unique to the individual and should not be correlated with other individual characteristics. Each entity is different therefore the entity’s error term and the constant (which captures individual characteristics) should not be correlated with the others. If the error terms are correlated then FE is no suitable since inferences may not be correct and you need to model that relationship (probably using random-effects), this is the main rationale for the Hausman test (presented later on in this document).

The equation for the fixed effects model becomes:

 (1)

Where  is the dependent variable (*DV*) where *i*= entity and *t*= time.

  is the coefficient for that independent variable,

  (*i*=1….*n*) is the unknown intercept for each entity (*n* entity-specific intercepts).

  is independent variable(s)

  is the error term

The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics.

“In the case of time-series cross-sectional data the interpretation of the beta coefficients would be “…for a given country, as *X* varies *across time* by one unit, *Y* increases or decreases by *β* units” (Bartels, Brandom, “Beyond “Fixed Versus Random Effects”: A framework for improving substantive and statistical analysis of panel, time-series cross-sectional, and multilevel data”, Stony Brook University, Is the coefficient for the binary repressors (entities) working paper, 2008.

Fixed-effects will not work well with data for which within-cluster variation is minimal or for slow changing variables over time.

Another way to see the fixed effects model is by using binary variables. So the equation for the fixed effects model becomes:

  (2)

Where  is the dependent variable (*DV*) where *i*= entity and *t*= time.

  is the coefficient for that independent variable,

 is the entity *n.* Since they are binary (dummies) you have n-1 entities included in the model.

  is independent variables

  is the coefficient for the binary repressors (entities)

  is the error term

Both eq.1 and eq.2 are equivalents:

“the slope coefficient on *X*is the same from one [entity] to the next. The [entity]-specific intercepts in [eq.1] and the binary regressors in [eq.2] have the same source: the unobserved variable *Z*ithat varies across states but not over time.” (Stock and Watson, 2003, p.280)

You could add time effects to the entity effects model to have a *time and entity fixed effects regression model*:

 (3)

Where  is the dependent variable (*DV*) where *i*= entity and *t*= time.

  is the coefficient for that independent variable,

 is the entity *n.* Since they are binary (dummies) you have *n-1* entities included in the model.

  is independent variables

  is the coefficient for the binary regressors (entities)

  is the error term

is time as binary variable (dummy), so we have *t-1* time periods.

is the coefficient for the binary time regressors.

Control for time effects whenever unexpected variation or special events my affect the outcome variable.

The fixed-effects model controls for all time-invariant differences between the individuals, so the estimated coefficients of the fixed-effects models cannot be biased because of omitted time-invariant characteristics…[like culture, religion, gender, race, etc]

One side effect of the features of fixed-effects models is that they cannot be used to investigate time-invariant causes of the dependent variables. Technically, time-invariant characteristics of the individuals are perfectly collinear with the person [or entity] dummies. Substantively, fixed-effects models are designed to study the causes of changes within a person [or entity]. A time-invariant characteristic cannot cause such a change, because it is constant for each person.”

**2. RANDOM-EFFECTSMODEL**

The rationale behind random effects model is that, unlike the fixed effects model, the variation across entities is assumed to be random and uncorrelated with the independent variables included in the model:

*“…the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not” [Green, 2008, p.183]*

If you have reason to believe that differences across entities have some influence on your dependent variable then you should use random effects. An advantage of random effects is that you can include time invariant variables (i.e. gender). In the fixed effects model these variables are absorbed by the intercept. The random effects model is:

 (4)

Where  is the dependent variable (*DV*) where *i*= entity and *t*= time.

  is the coefficient for that independent variable,

  (*i*=1….*n*) is the unknown intercept for each entity (*n* entity-specific intercepts).

  is independent variable(s)

  is the between entity-error term

  is the within entity-error

Random effects assume that the entity’s error term is not correlated with the predictors which allows for time-invariant variables to play a role as explanatory variables. In random-effects you need to specify those individual characteristics that may or may not influence the predictor variables. The problem with this is that some variables may not be available therefore leading to omitted variable bias in the model.

RE allows to generalize the inferences beyond the sample used in the model.

**FIXED OR RANDOM?**

**Hausman test**

To decide between fixed or random effects you can run a Hausman test where the null hypothesis is that the preferred model is random effects vs. the alternative the fixed effects (see Green, 2008, chapter 9). It basically tests whether the unique errors (*ui*) are correlated with the regressors, the null hypothesis is they are not. Run a fixed effects model and save the estimates, then run a random model and save the estimates, then perform the test.

xtreg*y x1*,fe

estimates store *fixed*

xtregy *x1*, re

estimates store *random*

hausman *fixed random*

Testing for random effects: Breusch-Pagan Lagrange multiplier (LM). The null hypothesis in the LM test is that variances across entities is zero. This is, no significant difference across units. The command in Stata is xttset0type it right after running the random effects model.

xtreg*y x1*,

rexttest0

Test: Var(u) = 0

chi2(1) = 2.67

Prob > chi2 = 0.1023

Here we failed to reject the null and conclude that random effects is not appropriate. This is, not significant differences across countries is found.

**Testing for time-fixed effects**

To see if time fixed effects are needed when running a FE model use the command testparm. It is a joint test to see if the dummies for all years are equal to 0, if they are then no time fixed effects are needed.

F( 9, 53) = 1.21

Prob > F = 0.3094

We failed to reject the null that all years coefficients are jointly equal to zero therefore no time fixed-effects are needed.

**Testing for fixed effects**

To see if fixed effects are needed use the command test. It is a joint test to see if the dummies for all entities are equal to 0, if they are then no fixed effects are needed.

 F( 6, 62) = 2.97

 Prob > F = 0.0131

We reject the null that all entities coefficients are jointly equal to zero therefore fixed-effects are needed (p-value <0.05)

**Testing for cross-sectional dependence/contemporaneous correlation**

Pasaran CD (cross-sectional dependence) test is used to test whether the residuals are correlated across entities\*. Cross-sectional dependence can lead to bias in tests results (also called contemporaneous correlation). The null hypothesis is that residuals are not correlated. The command for the test is xtcsd (which you have to install typing ssc install xtcsd)

xtreg*y x1*, fe

xtcsd, pesaran abs

Pesaran's test of cross sectional independence = 1.155, Pr = 0.2479 No cross-sectional dependence.

Had cross-sectional dependence be present Hoechle suggests to use Driscoll and Kraay standard errors using the command xtscc(install it by typing sscinstall xtscc). Type help xtscc for more details.

**Testing for heteroskedasticity**

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

xtreg y x1, fe

xttest3

chi2 (7) = 42.77

Prob>chi2 = 0.0000 Presence of heteroskedasticity

The null is homoskedasticity (or constant variance). Above we reject the null and conclude heteroskedasticity. To install the command xtest3 type ssc install xtest3. Type help xtest3 for more details.

**NOTE:** Use the option ‘robust’ to control for heteroskedasticiy (in both fixed and random effects).

**Testing for serial correlation**

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

Wooldridge test for autocorrelation in panel data.

: no first-order autocorrelation

F( 1, 6) = 0.214

Prob > F = 0.6603 No serial correlation

The null is no serial correlation. Above we fail to reject the null and conclude the data does not have first-order autocorrelation. To install the command xtserial type ssc install xtserial. Type help xtserial for more details.