Lecture 23. Panel Data Models.

Introduction to panel data analysis

Panel (or longitudinal) data set contains observations on the same units for several periods of time. Units can be individuals, countries, households, etc. Panel data sets are very typical in economic and social analysis, - like macroeconomic indicators for the regions of Russia in different time periods, or expenditures structure of households in different time units, or exam grades of ICEF students in different exam sessions, etc. In many countries there are regular surveys like Russian Longitudinal Monitoring Survey (RLMS), or National Longitudinal Survey of Youth (NLSY) in the USA, etc., which provide reliable panel data sets for all kinds of research and analysis.

Panel data set is **balanced** if every unit is surveyed in every time period and **unbalanced** otherwise. **The RLMS and NLSY present** unbalanced panel data, since some respondents can move, refuse, die, etc.

Benefits of having panel data

1. **Larger data set**. If in time series data T observations are available, with panel data *T*n* observations are available, where n is the number of units and T is the number of periods.
2. **Unobserved heterogeneity** problem (more about it later) is eliminated or mitigated.
3. **Dynamics** can be explored better (compared to cross-section). Although with cross-section one could investigate dynamics by asking retrospective questions, it is not very reliable, as people forget details over time.
4. Panel data are often of **higher quality**. For example, national surveys are usually rather well designed and well organised.

Panel Data Models

When dealing with the panel data, we may assume the following type of the DGP (data generating process):

$$Y_{i\,t} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{i\,t} \qquad (1)$$

where *X*s are observed and *Z*s are unobserved variables, *i* – unit index, *t* – time index, *j*, *p* – summation indices for observed and unobserved variables respectively. The term δt allows

intercept to shift over time. It can be substituted for dummies, representing corresponding periods if the constant change assumption is too strong.

Note: unobserved variables are assumed to be static, i.e. do not change over time, thus Z has no time index. As there is no information on Z, we can rename the variables:

$$\sum_{p=1}^{s} \gamma_p Z_{pi} = \alpha_i$$

α_i is referred to as **unobserved heterogeneity term**. It is also assumed that the disturbance term ε_{it} satisfies all the Gauss-Markov conditions, in particular that $\mathrm{Cov}(\alpha_i, \varepsilon_{jt}) = 0 \ , \forall\ i, j, t$

If α_i is ignored, by considering the model

$$Y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \delta t + \varepsilon_{it} \ ,$$

then OLS estimators will suffer from omitted variable bias, i.e., will be biased and inconsistent if α_i is correlated with any of the regressors. Thus, under such circumstances, α_i has to be accounted for somehow, and the Panel Data models allow to do it. Note that it is impossible for the cross section data.

Fixed effect model

The so-called Fixed Effect approach allows to transform the model (1) in such a way that the unobserved heterogeneity term is removed. Three different fixed effect type methods of transformation and estimation of the model (1) will be considered: First Differences Method, Within Groups Method, and Least Squares Dummy Variables (LSDV) Method.

First differences method is as follows:

We lag the model (1) by one period:

$$Y_{it-1} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ijt-1} + \alpha_i + \delta(t-1) + \varepsilon_{it-1} \ ,$$

And then subtract Y_{it-1} from Y_{it}:

$$Y_{it} - Y_{it-1} = \sum_{j=1}^{k} \beta_j (X_{ijt} - X_{ijt-1}) + \delta + \varepsilon_{it} - \varepsilon_{it-1} \ , \text{ or}$$

$$\Delta Y_{it} = \sum_{j=1}^{k} \beta_j \Delta X_{ijt} + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

Thus, the unobserved heterogeneity term α_i disappears. However, autocorrelation of the type MA(1) arises. In general, it leads to inefficiency of the OLS estimators and invalid test statistics, though consistency is provided. Only in the special case when the original disturbance term was subject to AR(1) autocorrelation with coefficient close to 1, then the disturbance term in the transformed model will not be autocorrelated.

Least squares dummy variables (LSDV) method

A natural way to account for α_i is to introduce a set of dummies for units. The model looks as follows:

$$Y_{it} = \sum_{j=1}^{k} \beta_j X_{jit} + \sum_{i=1}^{n} A_i D_i + \delta t + \varepsilon_{it}$$

where D_i is the dummy variable equal to 1 for i-th unit and zero otherwise. Thus, there are as many dummy variables as there are units, i.e. n. But note that such specification is possible only if intercept is omitted, otherwise one falls into the dummy trap. Alternatively, you can choose a reference category and keep the intercept.

Within groups method

This method allows to cancel the unobserved heterogeneity term by using deviations of the variables from their mean values. To apply it, for the model

$$(1)\ Y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it},$$

consider the average Y_i in time:

$$(2)\ \bar{Y}_i = \beta_0 + \sum_{j=1}^{k} \beta_j \bar{X}_{ij} + \alpha_i + \delta \bar{t} + \bar{\varepsilon}_i$$

As α_i is constant in time for each unit, it is not affected.

Them subtract (2) from (1):

(1) – (2)

$$Y_{it} - \bar{Y}_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it} - \beta_0 - \sum_{j=1}^{k} \beta_j \bar{X}_{ij} - \alpha_i - \delta \bar{t} - \bar{\varepsilon}_i$$

$$Y_{it} - \bar{Y}_i = \sum_{j=1}^{k} \beta_j \left(X_{ijt} - \bar{X}_{ij}\right) + \delta(t - \bar{t}) + \varepsilon_{it} - \bar{\varepsilon}_i$$

Thus the unobserved heterogeneity term α_i disappears, as well as the intercept β_0. This method is called «within groups» one because the variations of the dependent variable around its mean are regressed on the variations of explanatory variables around their means.

Dummy variables vs Within groups

It can be shown mathematically that within-groups method is equivalent to LSDV method. Thus, the two methods always give identical estimates. If the number of units is large, it is not convenient to introduce dummy variables. For this reason in practice within-groups method is used.

The only thing that can be unclear is degrees of freedom. In LSDV method there are $n*T–k-n$ degrees of freedom ($n*T$ observations, k regressors, and n dummies). At first glance, it looks as if within-groups method has $n*T–k$ degrees of freedom. Nevertheless, transformation of the model consumes n degrees of freedom through the calculation of

averages, so the number of the degrees of freedom is, as expected, the same for both methods.

So the Fixed Effect Model allows to take into account some unobservable heterogeneity. At the same time, the approach has essential drawbacks.

The drawbacks of the Fixed Effect model are:

1. All variables that are constant in time (though different for different units) disappear and we cannot determine to what extent they influence the dependent variable.
2. Variation of the new explanatory variables is (most likely) smaller than in the original specification. Thus, the precision of the estimates of the coefficients decreases. If measurement error bias took place, it would be aggravated.
3. n (number of units) degrees of freedom are lost as a result of the model transformation or of adding extra dummies.

Random effect model

If α_i is not correlated with regressors then it can be made a part of the disturbance term, and the new disturbance term will be uncorrelated with the regressors too. If so, omitting the individual effect will **not** make OLS estimators unbiased and inconsistent, while doing so will enable to save the degrees of freedom and keep regressors which are constant in time. The method allowing to do this is called a Random Effect Method.

Assumptions on the α_i in the Random Effect Model:

- α_i is taken randomly from a fixed distribution
- α_i is independent from the regressors

Thus, it is possible to move fixed part of α_i to the constant and the rest – to the disturbance term. Hence, without loss of generality, we will assume that $E(\alpha_i) = 0$. Then:

$$Y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \delta t + u_{it} ,$$

where $u_{it} = \alpha_i + \varepsilon_{it}$.

Assume that ε_{it} satisfies all Gauss-Markov conditions.

Let us check whether the new disturbance term u_{it} satisfies Gauss-Markov conditions:

1. $E(u_{it}) = E(\alpha_i + \varepsilon_{it}) = E(\alpha_i) + E(\varepsilon_{it}) = 0 + 0 = 0$
2. $Var(u_{it}) = Var(\alpha_i + \varepsilon_{it}) = Var(\alpha_i) + Var(\varepsilon_{it}) + 2Cov(\alpha_i,\varepsilon_{it})$=[assume $Cov(\alpha_i,\varepsilon_{it})$= =0] = $Var(\alpha_i) + Var(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$
3. $Cov(u_{it}, u_{it-s}) = Cov(\alpha_i + \varepsilon_{it}, \alpha_i + \varepsilon_{it-s})$ =[assume ε_{it} is not serially correlated]= =$Var(\alpha_i) = \sigma_\alpha^2$

4. $\text{Cov}(X_{ijt}, u_{it}) = \text{Cov}(X_{ijt}, \alpha_i + \varepsilon_{it}) = \text{Cov}(X_{ijt}, \alpha_i) + \text{Cov}(X_{ijt}, \varepsilon_{it}) = 0$

Thus, the Gauss-Markov condition 3 is violated. The disturbance term is autocorrelated for each unit *i,* with a particular type of autocorrelation. To overcome this problem, the so-called **Feasible Generalised Least Squares (FGLS)** method is used. This method takes autocorrelation into account and produces the best linear unbiased and consistent estimators, possible with the available information.

Random vs Fixed

Random effect model is superior to Fixed effect model for two reasons:

1. n degrees of freedom are not lost in the Random effect model
2. observations that are constant in time are not dropped

So, it is better to use Random Effect than Fixed Effect model. However, if at least one of the assumptions of Random Effect model does not hold, we have to use Fixed Effect model. Thus, the question is whether the necessary assumptions hold. As for the first assumption (randomly drawn observations), it is supposed to be guaranteed by the survey. The tests for checking the assumption on independence of the new disturbance term with the explanatory variables, are described in the section "Tests".

Pooled regression

It can be the case that there is no unobserved heterogeneity at all, i.e.

$$\alpha_i = 0, \forall\, i$$

If so, the sample can be pooled (unified sample for all units and periods). This will give two benefits comparing to the Random effect model since there is no need to allow for non-existing individual effects:

- it allows to get not just unbiased and consistent but also efficient estimates, with valid tests, without special procedures like FGLS, and these properties are provided for finite samples;
- no loss of efficiency due to no testing of irrelevant constraints or effects.

Panel data summary

True situation	Fixed Effect Model	Random Effect Model	Pooled Regression
$\alpha_i = 0$ for any *i*	• unbiased • consistent	• unbiased • consistent	• unbiased • consistent

	• inefficient	• inefficient	• efficient
$cov(\alpha_i, X_j) = 0$ for any *i,j*	• unbiased • consistent • inefficient	• unbiased • consistent • efficient	• unbiased • consistent • inefficient
$cov(\alpha_i, X_j) \neq 0$ for some *i,j*	• unbiased • consistent • efficient	• biased • inconsistent	• biased • inconsistent

Tests

In this section tests relevant to panel data models are described. There are two main questions that one can ask on the model to use:

1. Should we use Random effect or Fixed effect model?
2. Should we use Pooled regression or Random Effect model (if Random effect is preferred to Fixed effect in the p.1)?

For the p.1, one can use the Durbin-Wu-Hausman (DWH) test, while Breush-Pagan Lagrange Multiplier test can be used for the p.2.

Durbin-Wu-Hausman test

This test is applied in many cases other than panel data, e.g. to detect the endogeneity or measurement errors problems. Its purpose is to test whether there is significant difference between the estimates of coefficients, obtained by different methods, one set always consistent while another consistent only under particular circumstances (some hypothesis H_0 implementation).

$$Y_{i\,t} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{i\,t}$$

To do the test, we have to consider first whether all α_i can be considered as random variables taken from the same distribution. Under H_0 (which means that α_i are not correlated with X_j for any *i,j*) both Fixed Effect and Random Effect models provide us with consistent estimates. However Fixed Effect model estimates will be inefficient since it involves estimating an unnecessary set of coefficients, so **Random Effect** model should be used if H_0 is not rejected.

If we reject H_0 in favour of H_a (meaning thet α_i and X_j are correlated for some *i,j*), the Random Effect estimators would be biased and inconsistent, while **Fixed Effect** estimators are consistent, and hence the **Fixed Effect model should be used** though having the drawbacks indicated above.

Under H_0 the DWH test statistic has a chi-square (χ^2) distribution. Its calculations involve matrix algebra. The number of degrees of freedom usually equals the number of coefficients compared, but can also be lower in some special cases.

Random Effect model or Pooled Regression?

In order to decide if we should use the panel data models at all, we have to test if there is unobserved heterogeneity. For this, Breush-Pagan Lagrange multiplier test could be used. The null hypothesis H_0 for it means that $\alpha_i = 0$ for any *i*. The test statistic under H_0 has χ^2 distribution with one degree of freedom.

If we do not reject the H_0 then Pooled regression should be estimated, and it provides unbiased, consistent and efficient estimates. If H_0 is rejected then we apply the Random Effect model.