Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 9
Panel Data Analysis

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Panel Data

Panel Data or Longitudinal Data are multi-dimensional data involving measurements over time. Panel data contain observations of multiple phenomena obtained over multiple time periods for the same objects. Panel data have three dimensions:

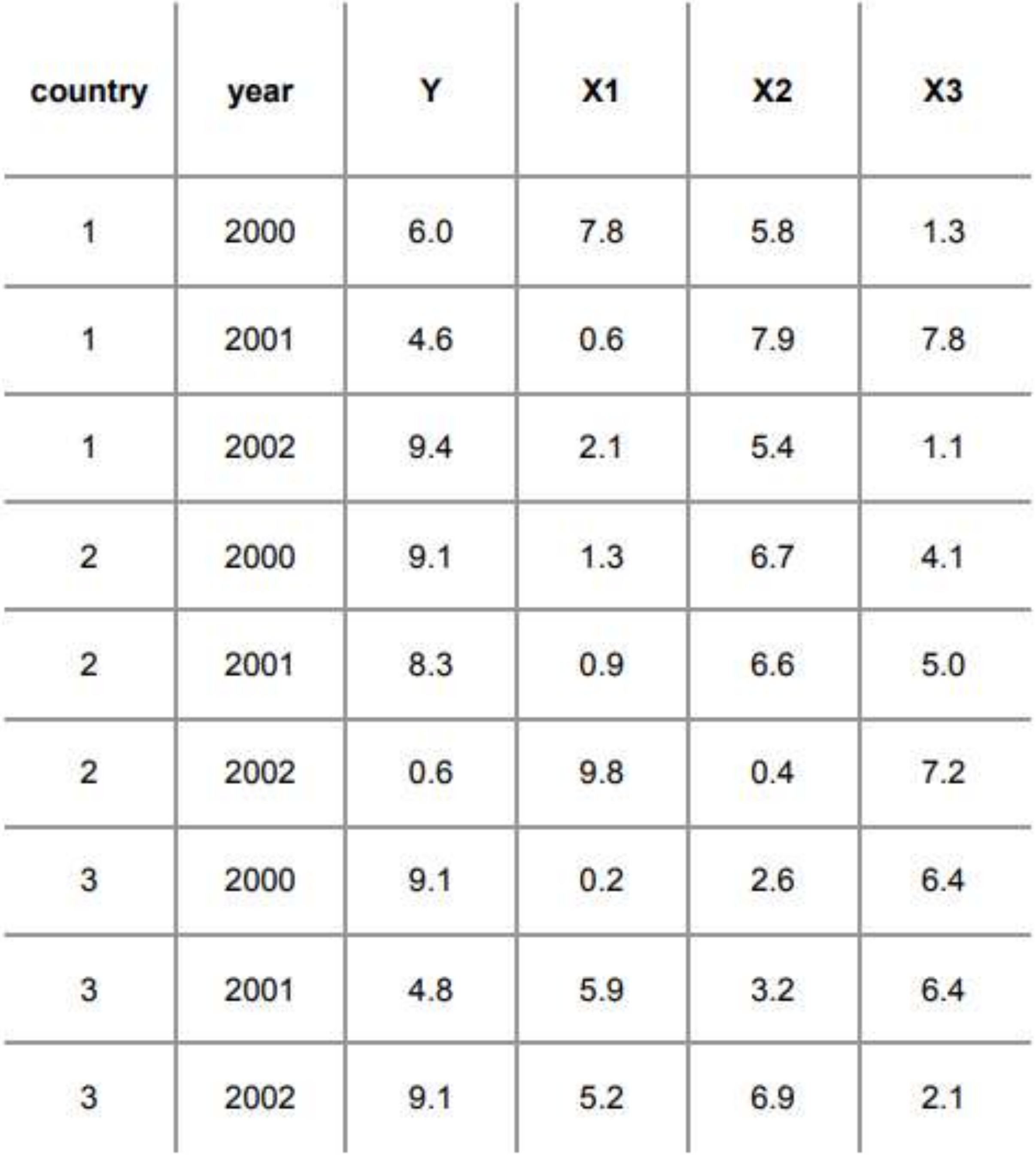
1) variables,

2) cases,

3) time.

Example of panel data

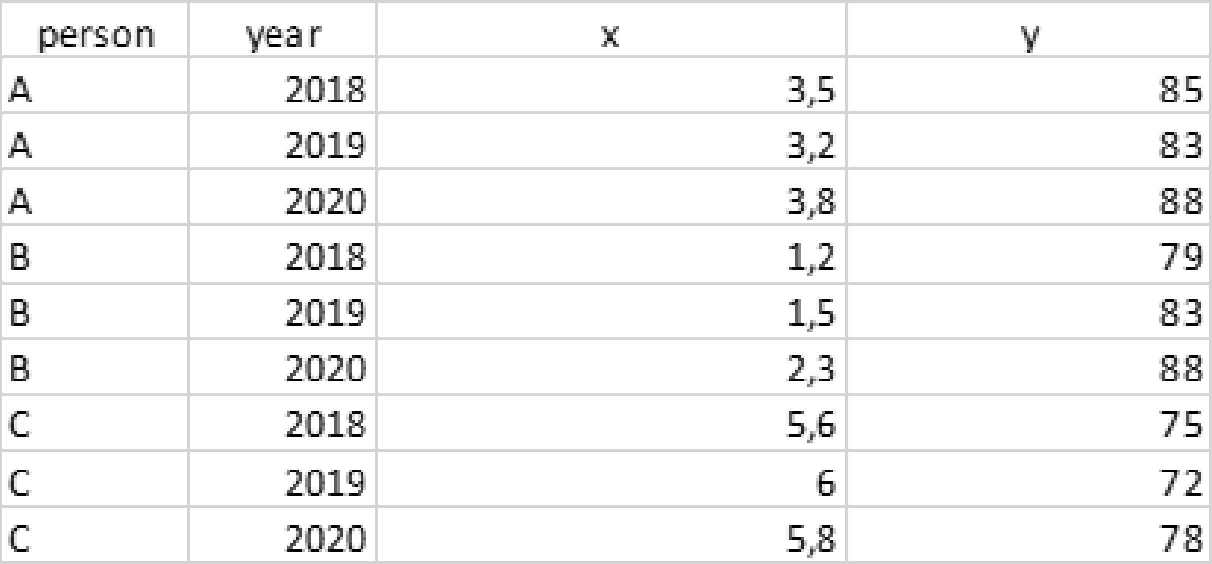
In general, panel data can be seen as a combination of cross-sectional and time-series data. Cross-sectional data is described as one observation of multiple objects and corresponding variables at a specific point in time (i.e. an observation is taken once). Time-series data only observes one object recurrently over time. Panel data comprises characteristics of both into one model by collecting data from multiple, same objects over time. In a nutshell, we can think of it like a timeline in which we periodically observe the same objects.



country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Example of panel data

It is important to note that we always need one column to identify the indiviuums under observation (column *person*) and one column to document the points in time the data was collected (column *year*). Those two columns should be seen as **multi-index**.



person	year	x	y
A	2018	3,5	85
A	2019	3,2	83
A	2020	3,8	88
B	2018	1,2	79
B	2019	1,5	83
B	2020	2,3	88
C	2018	5,6	75
C	2019	6	72
C	2020	5,8	78

Sample Panel Dataset

Panel Data Advantages

Due to a special structure, panel data allows to build more flexible and meaningful models and receive answers to questions that are not available in the framework of models based on two-dimensional datasets. They allow to consider and analyze **individual differences between** units of analysis (to solve the problem of bias in the results obtained due to unobservable data heterogeneity).

For example, in cross-country data in a sample of 100 countries their **individual characteristics** may influence the results of the assessment. When working with two-dimensional datasets, it is difficult to take these individual features into account. Analysis of panel data allows them to be considered and to avoid biasing results that would be if these individual features were ignored.

Panel Data Advantages

Another advantage – we can explore the dynamics for several objects at once. That is, not a change in GDP in Russia, as in the analysis of time series, but a change in GDP for several countries at once. When working with panel data, there is more data (observations) than when working with two-dimensional samples and time series. And this allows you to get more accurate results.

Balanced Panels

Balanced panel – for each object there is an observation for each moment in time, i.e., number of observations = n* T (there are no missing values within the dataset).

Unbalanced panel - there are gaps in the data, i.e., number of observations < n* T.

If the occurrence of gaps is exogenous, that is, random and does not correlate with the dependent variable, then for unbalanced panels you can use the same analysis methods as for balanced ones.

Types of regressors

Varying: vary in time and differ among different sample objects, for example, annual income, consumption.

Unchanged in time (time-invariant): gender, race, level of education.

Individual-invariant: vary in time, but do not differ for different sample objects, for example, the unemployment rate.

Example

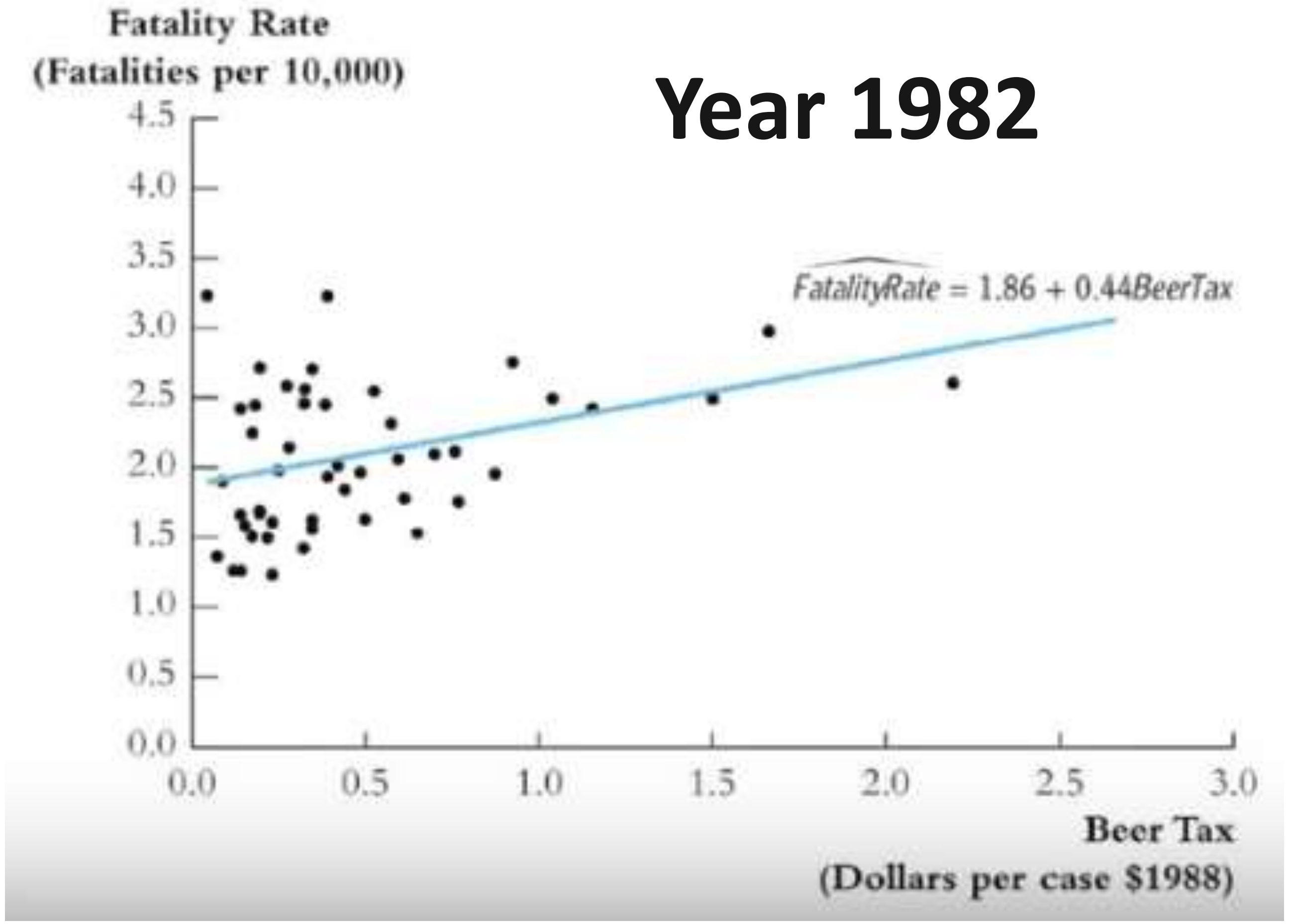
Research question: How are beer taxes related to road traffic fatality rates?

Data: 48 US states for 1982 and 1988.

Variables:
Fatality Rate – the number of deaths from road accidents per year per 10 thousand inhabitants;
Beer Tax – tax on the sale of beer.

Hypothesis: the relationship between these variables is negative.

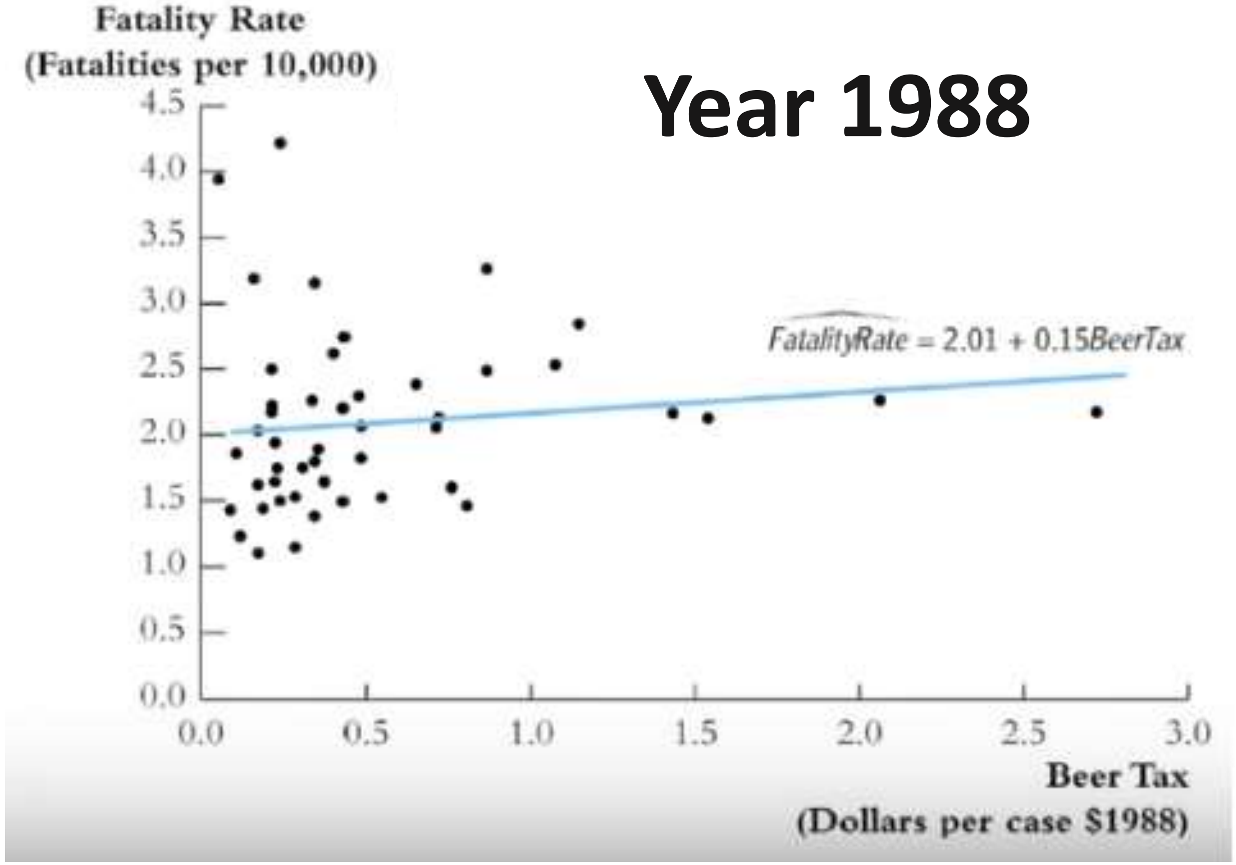
Example



Year 1982

We can see a positive relationship on the graph. Such unexpected result may be due to the fact that we incorrectly evaluate the data or do not take into account something important.

Example



Year 1988

Possible explanation for this and the previous graphs: fatality rate not only depends on the beer tax, but also on some individual characteristics of each state.

Example

$$\begin{gathered} FatalityRate_{it} = \\ = \beta_0 + \delta * t + \beta_1 * BeerTax_{it} + \gamma * \boxed{Z_i} + \varepsilon_{it} \end{gathered}$$

To take into account the individual characteristics of each state we add Zi into the equation. The difficulty with the variable Zi is that it is not observable. It is very difficult to measure the cultural characteristics of the states and directly take them into account in the model.

Example

Let’s take two equations for 1982 and 1988.

For 1982 (t=1):

$$FatalityRate_{i,1} = \\ = \beta_0 + \delta * 1 + \beta_1 * BeerTax_{i,1} + \gamma * Z_i + \varepsilon_{i,1}$$

For 1988 (t=2):

$$FatalityRate_{i,2} = \\ = \beta_0 + \delta * 2 + \beta_1 * BeerTax_{i,2} + \gamma * Z_i + \varepsilon_{i,2}$$

If we subtract the first equation from the second equation Zi **will disappear** =>

Example

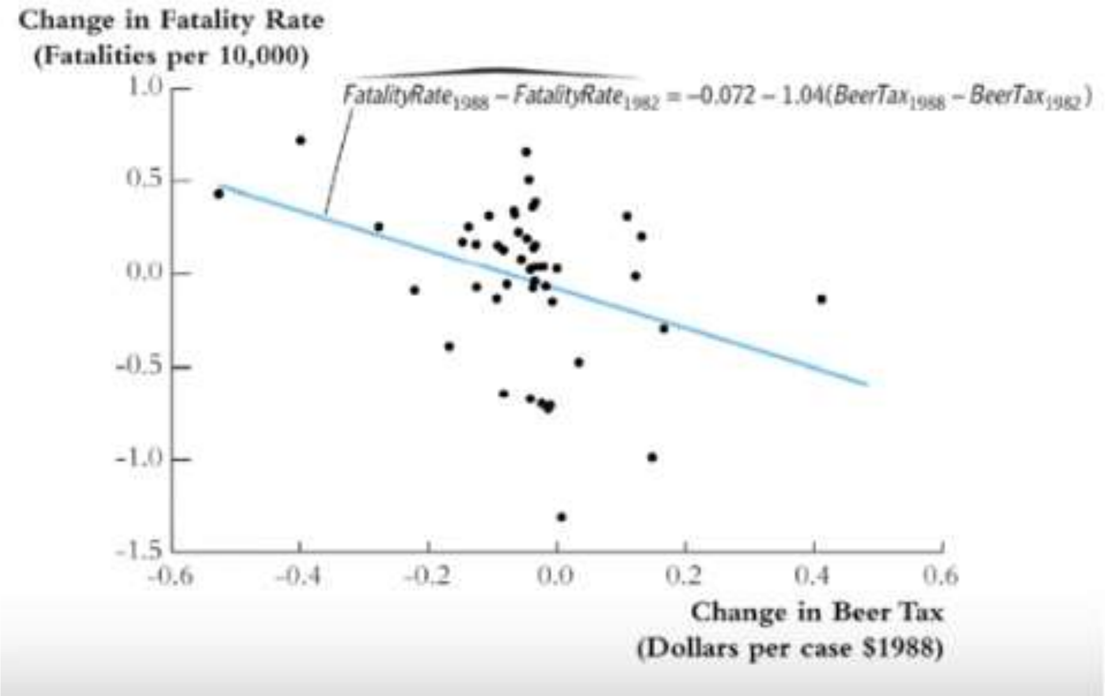
$$FatalityRate_{i,2} - FatalityRate_{i,1} =$$
$$= \delta + \beta_1 * \left(BeerTax_{i,2} - BeerTax_{i,1}\right) + u_i$$

$$u_i = \varepsilon_{i2} - \varepsilon_{i1}$$

$$\Delta FatalityRate_i = \delta + \beta_1 * \Delta BeerTax_i + u_i$$

With this subtraction, the variable ***Zi*** **disappeared**. So, if we evaluate how the **change in y depends on the change in x**, then the individual characteristics of each state will disappear in this equation. The unobservable variable for which we have no statistics will be removed from the equation. Since it is no longer in the equation, there will be no bias due to the omission of an essential variable.

Example



This graph shows the relationship between the **change in fatality rate and the change in beer tax**. Dependence became negative. The described approach to the estimation of regression on panel data is called the model with the first differences. This is one of the possible approaches.

Panel data models

1) Pooled model;
2) Fixed effects model:
    - First difference model,
    - Least squares dummy variables model,
    - Within groups fixed effects.
3) Random effects model.

Pooled model

Pooled model can be described as simple regression model that is performed on panel data. It ignores time and individual characteristics and focuses only on dependencies **between** the individuums.

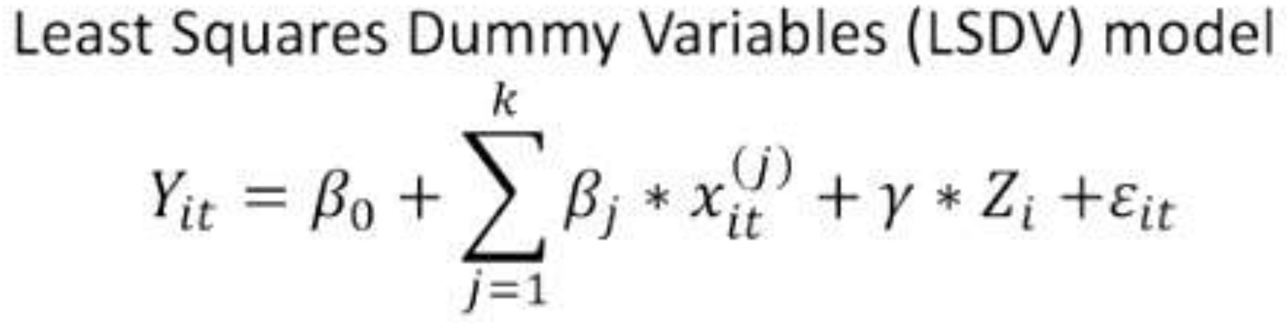
Fixed-Effects (FE) Model

FE-model determines individual effects of unobserved, independent variables as constant (“fix“) over time.

Random-Effects (RE) Model

RE-models determine individual effects of unobserved, independent variables as random variables over time.

Least Squares Dummy Variables model



Least Squares Dummy Variables (LSDV) model

$$Y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j * x_{it}^{(j)} + \gamma * Z_i + \varepsilon_{it}$$

In this approach, instead of using the first differences, we can add **several binary variables** to the model. Thus, we get many dummy variables. One variable per each object (state).

Least Squares Dummy Variables model

$$Y_{it} = \sum_{j=1}^{k} \beta_j * x_{it}^{(j)} + \alpha_1 * D1_i + \cdots \alpha_n * Dn_i + \varepsilon_{it}$$

$$\alpha_1 = \beta_0 + \gamma * Z_1 \quad \ldots \quad \alpha_n = \beta_0 + \gamma * Z_n$$

If we add them to the equation together with the regressors, we get a new equation, in which there will be all the original regressors and **dummy variables, each of which is responsible for its own object**, for example, for its own state. And these variables will allow us to take into account the individual characteristics of the objects of study.

Least Squares Dummy Variables Model

In the first difference model, we simply **eliminated** individual effects from our model. We made transformations during which the unobservable variable, which is responsible for individual effects, **disappeared** from the equation. We received unbiased estimates of the coefficients in regressions, but at the same time we could not say anything about the individual features of each of the objects. In a model with dummy variables, we can **estimate** the coefficient for each dummy variable, and this coefficient will characterize the **individual features** for each of the objects.

Within group fixed effects

Initial equation:

$$y_{it} = \beta_0 + \beta_1 * x_{it} + \gamma * z_i + \varepsilon_{it}$$

Mean values for observed time periods:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \gamma * z_i + \bar{\varepsilon}_i$$

First minus second:

$$(y_{it} - \bar{y}_i) = \beta_1 * (x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$y_{it} = (y_{it} - y_i), \qquad x_{it} = (x_{it} - x_i),$$
$$\widetilde{\varepsilon_{it}} = (\varepsilon_{it} - \bar{\varepsilon}_i)$$

Result:

$$\widetilde{y_{it}} = \beta_1 * \widetilde{x_{it}} + \widetilde{\varepsilon_{it}}$$

This transformation is called within-group transformation. The resulting model is called the within-group regression. It explains the variation of the dependent variable around the average value for a group of observations related to a particular object.

Which model to select?

We should answer the question "Are there individual effects in the data?" If yes, we need to use a model with effects (fixed or random). If not, then pooled regression. To answer this question there is a special test that allows us to choose the right model.

Model with random effects

When using models with fixed effects, we believe that the individual effect for each of the objects that we model is a constant, a fixed value.

An alternative approach is the perception of individual effects as random variables. Then in our model there are two “sources of randomness”:

- random errors ε_i,
- random effects u_i (unchanged in time for a particular object).

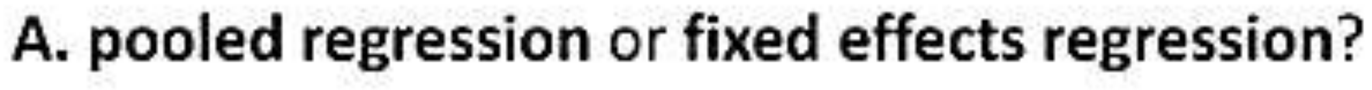
Choice between a model with random and fixed effects

The advantage of a model with random effects is that it can include regressors, which for each object are constant over time. But the limitation of this model is that regressors and individual effects should not correlate (this limitation is absent in the model with fixed effects).

It is recommended to use a random effects model when objects are randomly extracted from the general population. For example, we randomly select a thousand organizations out of a million. If we randomly sample again, we will select a different set of 1000 organizations with different set of individual effects. Then the individual effects will be random.

For country or regional samples, fixed effects models are usually used because a different set of countries or regions cannot be selected, it is fixed.

Select a model



A. pooled regression or fixed effects regression?

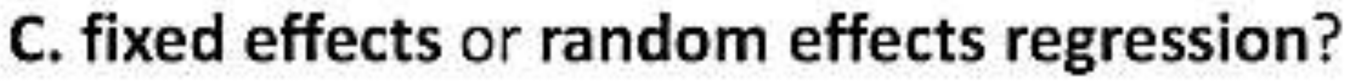
H0: all individual effects are the same and equal to zero.

If H0 is rejected the fixed effects regression should be selected.

B. pooled regression or random effects regression?

H0: the variance of individual effects is zero

If H0 is rejected the random effects regression should be selected.



C. fixed effects or random effects regression?

H0: individual effects and regressors are not correlated

If H0 is rejected the fixed effects regression should be selected.

Useful links

- http://www.cantab.net/users/bf100/pdf/pd_slides_fingleton.pdf
- https://bashtage.github.io/linearmodels/panel/examples/data-formats.html
- https://dspyt.com/panel-data-econometrics-an-introduction-with-an-example-in-python/
- https://bashtage.github.io/linearmodels/panel/examples/examples.html

Thank you for your attention!