MICROECONOMETRICS
CLASS 9

Wiktor Budziński

PANEL DATA

In some instances, our data can have a panel structure

- Several observations per one unit/individual
- Usually, observed over several subsequent time periods

It is likely that with such a format, error terms of the same individual will be correlated over time

- Individual-specific unobserved effects
- Some dynamic effects

A general approach would be to consider a T-variate distribution and estimate all correlations between error terms

- Usually too complicated

PANEL DATA

To simplify the issue, the oft-used specification is an error component model

$$\begin{cases} y_{it} = x_{it}\beta + v_{it} \\ v_{it} = u_i + \varepsilon_{it} \end{cases}$$

In this case, u_i measures individual-specific heterogeneity

- E.g., some bundle of the unobserved characteristics of the consumer

As such, the model assumes that correlation between any two error terms of the same individual is constant

- To account for it in the OLS one can use some robust variance-covariance matrices
  - Clustered / sandwich

PANEL DATA

If error component is independent from the observed variables OLS will still be consistent, but not efficient

- Basically, the assumption about spherical error terms is not met

Random effect model is more efficient solution which acknowledges the structure of the error terms

- Utilizes Generalized Least Squares method to obtain the estimates

RE model decomposes variances into individual-level and individual and time-level

PANEL DATA

One can try to get rid of the error-component by simply averaging the data over time

$$\bar{y}_i = \bar{x}_i \beta + u_i + \bar{\varepsilon}_i$$

Such between-estimator is generally not recommended though

- We lose a lot of information and work with significantly smaller sample
- Often not usable for the purpose of the study, for example difference-in-differences

PANEL DATA

The within-estimator, or fixed-effect model is probably the most widely used specification in the case of linear model

- You can think about is as having a dummy variable for each individual

This is not feasible in general, so usually one transforms the data

- Pooled OLS minus the between estimator
- This transformation get rids of the error component

$$\left(y_{it}-\overline{y}_i\right)=\left(x_{it}-\overline{x}_i\right)\beta+\left(u_i-u_i\right)+\left(\varepsilon_{it}-\overline{\varepsilon}_i\right)$$

- We can then estimate OLS on the transformed data
  - Note that there is still correlation between error terms, so some robust matrix is advised

PANEL DATA

The main advantage of the within-estimator is that it does not assume anything about the error component

- Specifically, error component can be correlated with the independent variables
- RE model and OLS will be biased in such a case, as it will basically lead to endogenous model

Within-estimator can be then considered one of the solutions for the endogeneity

- Does not require any instruments or indicators
- Requires multiple observations per individual
  - At least two
- Assumes that endogeneity is caused by individual-specific effect

PANEL DATA-TESTS

There are some tests that could be used to evaluate whether there should be an error component in our model

Breush-Pagan test uses OLS results and checks whether restriction of no random effect is justified

- It is a Lagrange Multiplier test, based on MLE estimator

F-test uses results from within-estimator

- Basically, compares residuals from within-regression with the residuals from OLS
- If there are not significantly different then there is no error component

PANEL DATA-TESTS

Hausman test can be used to evaluate whether we should be using RE or FE

- Test on whether the error component is correlated with independent variables
- Under the null we assume no correlation
  - RE will be consistent and efficient
  - FE will be consistent
- Under the alternative
  - RE will be inconsistent
  - FE still consistent

PANEL DATA-TESTS

The alternative test for the error component being of the fixed-effect type is the Chamberlain test, later updated by the Angrist and Newey

- We check whether residuals are a function of the independent variables for each period
- If they are jointly significant, then we can conclude that the fixed effect specification is appropriate
- Available only for the balanced panels

PANEL DATA

Often it is useful to add time-specific error component

$$\begin{cases} y_{it} = x_{it}\beta + v_{it} \\ v_{it} = u_i + \tau_t + \varepsilon_{it} \end{cases}$$

There exist some sort of transformations that can allow for estimation of individual- and time-specific effects with the within-estimator

As time dimension is usually low, it is often easier to just add dummy variables to the basic model

DYNAMIC PANEL DATA

The actual error structure could be sometimes more complex than in the error component model

Correlation between errors terms may not always be equal

- For example, correlation may get weaker with time

Most often this is modelled with the AR(1) process

$$\begin{cases} y_{it} = x_{it}\beta + v_{it} \\ v_{it} = u_i + \rho\varepsilon_{it-1} + \varepsilon_{it} \end{cases}$$

Can be tested for both FE and RE models

- Not straightforward to estimate in the case of FE
- For RE one can use ML

EXERCISE 1: PANEL DATA

1. Read *panelwages.xlsx* data into R
2. Compare the *pooled*, *between*, *within*, and *random* specifications of the panel model
    1. Test for the significance of the error component
    2. Test whether the error component is correlated with independent variables
    3. Test for the serial autocorrelation of the error terms (AR(1) process)
3. Estimate a random effect model with serial autocorrelation
4. Compare different models in *Sim_examples8*.R under different DGP

DIFFERENCE-IN-DIFFERENCES

One of the applications of the panel data is the so-called Difference-in-differences regression

- This is one of the treatment effect methods that utilizes additional information from the panel structure

By estimating a fixed effect model one can control for potential endogeneity (selection on unobservables)

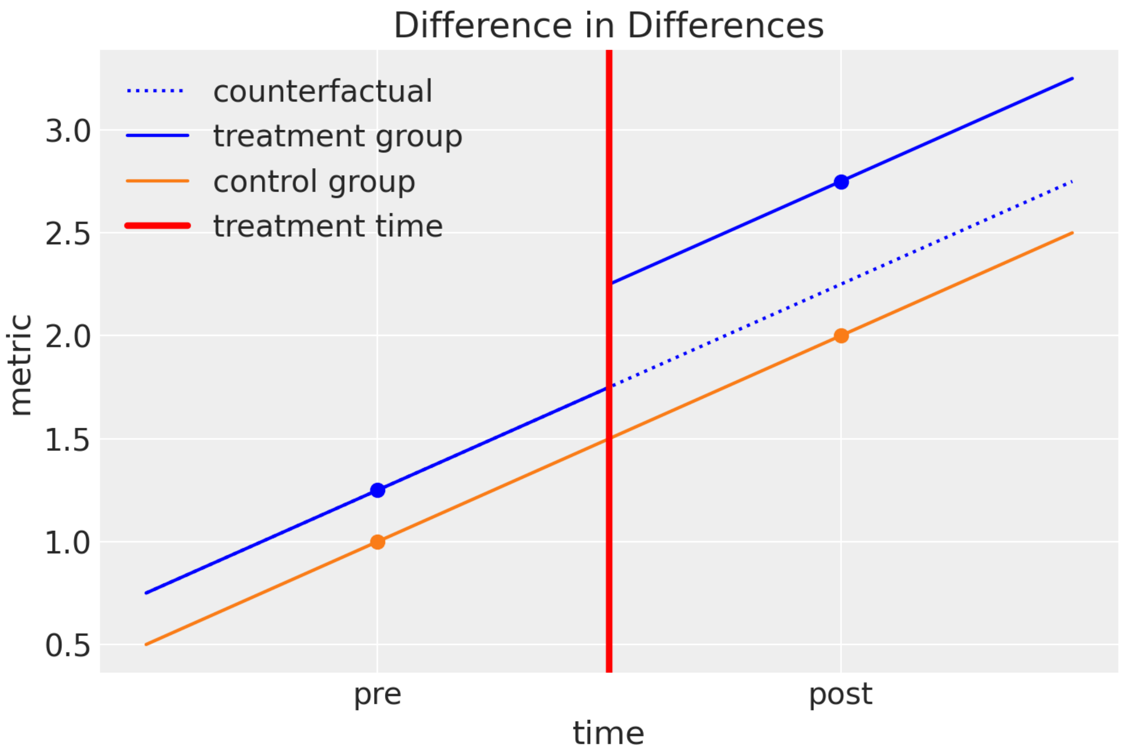
DiD can be used to study the effect of a certain event that takes place in-between two waves of the panel data

- The requirement is that we need to observe the control group and the treatment group before and after the event

DIFFERENCE-IN-DIFFERENCES

The crucial assumption of DiD is the <u>parallel trend assumption</u>

- If there was no event, then the difference between control and treated would stay the same



DIFFERENCE-IN-DIFFERENCES

In general, we are interested in estimating the model

$$y_{it} = E_{it}\alpha_1 + T_{it}\alpha_2 + E_{it}T_{it}\alpha_3 + x_{it}\beta + u_i + \varepsilon_{it}$$

Where *E* indicates whether the period is before or after the event

T indicates whether someone is in the treatment group or control group

Their interaction gives us the effect of the event

EXERCISE 2: DID

1. Read *napster.xlsx* data into R
2. Use DiD to evaluate the effect of *Napster* on people's expenditure on physical records
   - Peer-to-peer file sharing application, mostly used for sharing mp3 files over the internet

PANEL DATA FOR NONLINEAR MODELS

Estimating error components models in the case of non-linear models is not straightforward

- In most cases there is no within-estimator that could allow for estimation of fixed effects
- The solution could be to just add a separate dummy variable for each individual, but this will often lead to inconsistent estimates of betas
    - The incidental parameter problem
- For this reason, most non-linear models use a random effect specification to deal with the panel data
    - If we suspect that individual effects could be endogenous, then we have to use some other method to deal with it, for example, control function

In the case of non-linear models, if we ignore the error component, then the estimates will be biased and inconsistent

PANEL DATA FOR NONLINEAR MODELS

Poisson model is one of the exceptions for which the within-estimator exists

- It kind of exists for logit as well, but often requires dropping a large portion of the sample

Similarly as in the linear model, we can condition the model on some sufficient statistic from the data

- In the linear model we used a mean for each individual
- In the Poisson model we use the sum for each individual
  - This is strictly a characteristic of this particular distribution, caused by the fact that sum of Poisson variables is a Poisson variable

PANEL DATA FOR NONLINEAR MODELS

For Poisson model we specify the multiplicative error component:

$$\begin{cases} y_{it} \sim \mathrm{Poiss}\left(\eta_i \lambda_{it}\right) \\ \lambda_{it} = \exp\left(X_{it}\beta\right) \end{cases}$$

One can show that

$$P\left(y_i \mid Y_i, \eta_i\right) = \frac{Y_i\,!}{\Lambda_i^Y} \prod_t \left(\frac{\lambda_{it}^{y_{it}}}{y_{it}\,!} \right)$$

Above probability is not a function of the error component

$$Y_i = \sum_t y_{it}$$

$$\Lambda_i = \sum_t \lambda_{it}$$

PANEL DATA FOR NONLINEAR MODELS

If we assume a particular distribution for the error component, we will obtain a random effect model

- For Poisson distribution error component is usually specified as a Gamma distribution
- This leads to a relatively simple formula for probability, with one additional coefficient to estimate, which describes the Gamma distribution

$$P\left(y_i \mid \eta_i\right)=\frac{\delta^{\delta}\Gamma\left(Y_i+\delta\right)}{\Gamma\left(\delta\right)\left(\Lambda_i+\delta\right)^{Y_i+\delta}}\prod_t\left(\frac{\lambda_{it}^{y_{it}}}{y_{it}!}\right)$$

EXERCISE 3: POISSON PANEL DATA

1. Investigate the effect of hosting an article at ACTT (American Type Culture Collection) on its citations using the *GiantsShoulders* dataset
2. Formulate the model as a DiD-style regression
   1. Compare the fixed effect model with the random effect model

RANDOM PARAMETER MODELS

The direct extension of the random effect model is a random parameter model

Random effect model assumes that constant has some distribution in the population

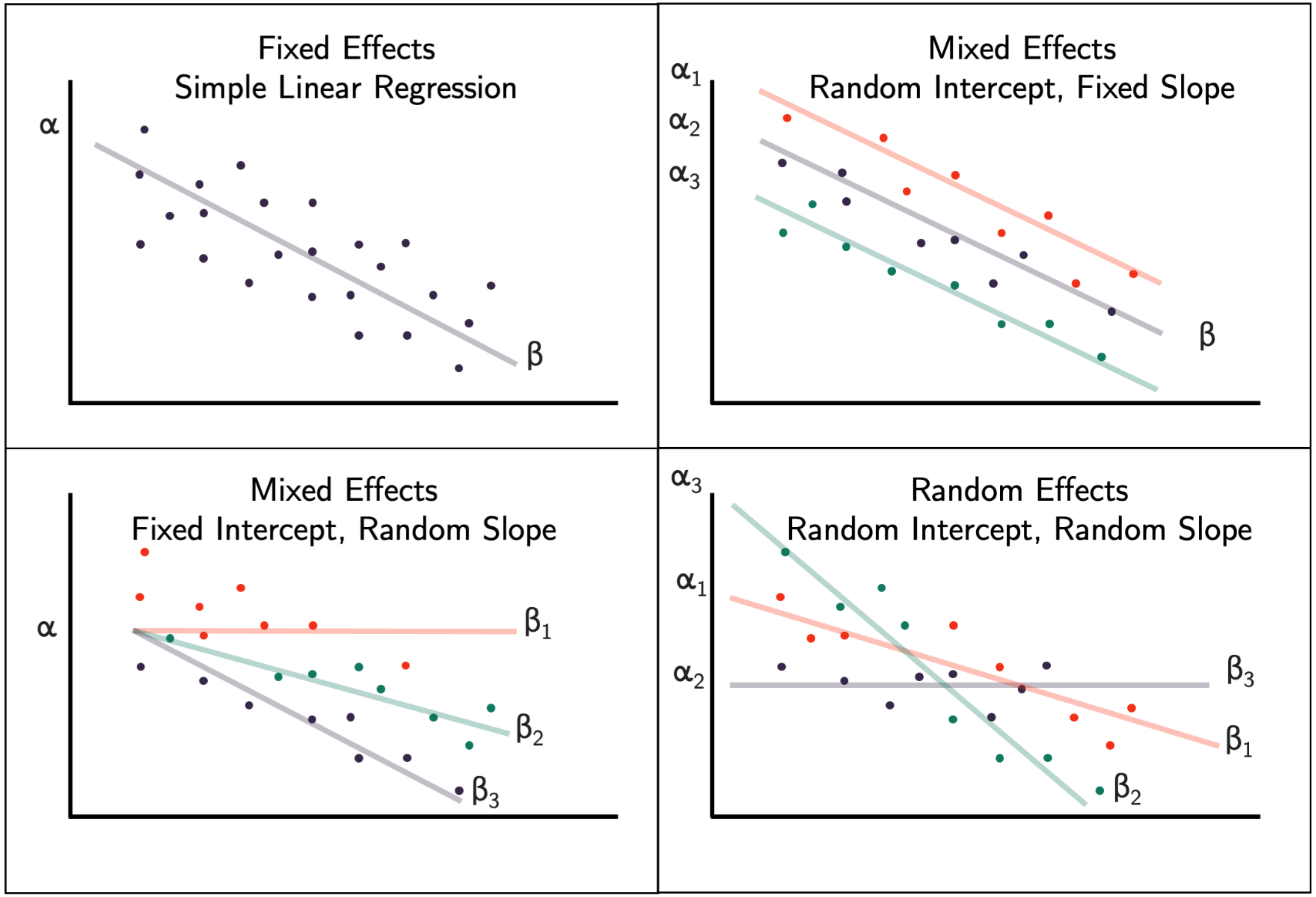
- Every individual have a different constant

Random parameter model extends it to all coefficients

- Every individual have a different set of parameters

As it is not feasible to estimate all of them, we need to assume some distribution for them

RANDOM PARAMETER MODELS



RANDOM PARAMETER MODELS

For example, for choice data

$$U_{ijt} = \boldsymbol{\beta}_i \mathbf{X}_{ijt} + \varepsilon_{ijt}$$

If we would know the values of these parameters, then choice probability would be given by by the product of MNL formulas

$$P\left(y_i \mid \boldsymbol{\beta}_i\right) = \prod_t \frac{\exp\left(\boldsymbol{\beta}_i \mathbf{X}_{ijt}\right)}{\sum_l \exp\left(\boldsymbol{\beta}_i \mathbf{X}_{ilt}\right)}$$

To calculate the unconditional probability we calculate the expected value over the random parameters

$$L_i = P\left(y_i\right) = \mathbf{E}\left[P\left(y_i \mid \boldsymbol{\beta}_i\right)\right] = \int P\left(y_i \mid \boldsymbol{\beta}_i\right) f\left(\boldsymbol{\beta}_i \mid \Omega\right) d\boldsymbol{\beta}_i$$

RANDOM PARAMETER MODELS

Solving multidimensional integrals is hard, so usually we simulate this expected value using Monte Carlo methods

- We take R draws from the distribution defined by $f\left(\boldsymbol{\beta}_i \mid \Omega\right)$, let denote them by $\boldsymbol{\beta}_i^r$
- We then simply take the average

$$L_i = P\left(y_i\right) = \mathbf{E}\left[P\left(y_i \mid \boldsymbol{\beta}_i\right)\right] \approx \frac{1}{R}\sum_{r=1}^{R} P\left(y_i \mid \boldsymbol{\beta}_i^r\right)$$

RANDOM PARAMETER MODELS

Examples:

- For the normal distribution we can generate draws from the standard normal distribution, $D_i^r \sim N(0,1)$ and then transform them with mean, and std. dev. coefficients:

$$\beta_i^r = \mu + \sigma D_i^r$$

- For the log-normal distribution we can use the same draws but with different transformation:

$$\beta_i^r = \exp\left(\mu + \sigma D_i^r\right)$$

RANDOM PARAMETER MODELS

Example:

- If we want to generate correlated draws we usually use multivariate normal distribution. Let say we want to have 3 correlated vectors of draws. We first generate three independent vectors of draws from standard normal distribution $D_{i1}^{r}, D_{i2}^{r}, D_{i3}^{r}, \sim N(0,1)$
- We can then create correlated draws in following way

$$\begin{cases} \beta_{i1}^{r} = \mu_1 + \sigma_1 D_{i1}^{r} \\ \beta_{i2}^{r} = \mu_2 + \rho_{12} D_{i1}^{r} + \sigma_2 D_{i2}^{r} \\ \beta_{i3}^{r} = \mu_3 + \rho_{13} D_{i1}^{r} + \rho_{23} D_{i2}^{r} + \sigma_3 D_{i3}^{r} \end{cases}$$

RANDOM PARAMETER MODELS

Example (continued):

- Written in vector form: $\boldsymbol{\beta}_i^r = \boldsymbol{\mu} + \Gamma \cdot \mathbf{D}_i^r$
- Where

$$\Gamma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ \rho_{12} & \sigma_2 & 0 \\ \rho_{13} & \rho_{23} & \sigma_3 \end{bmatrix}$$

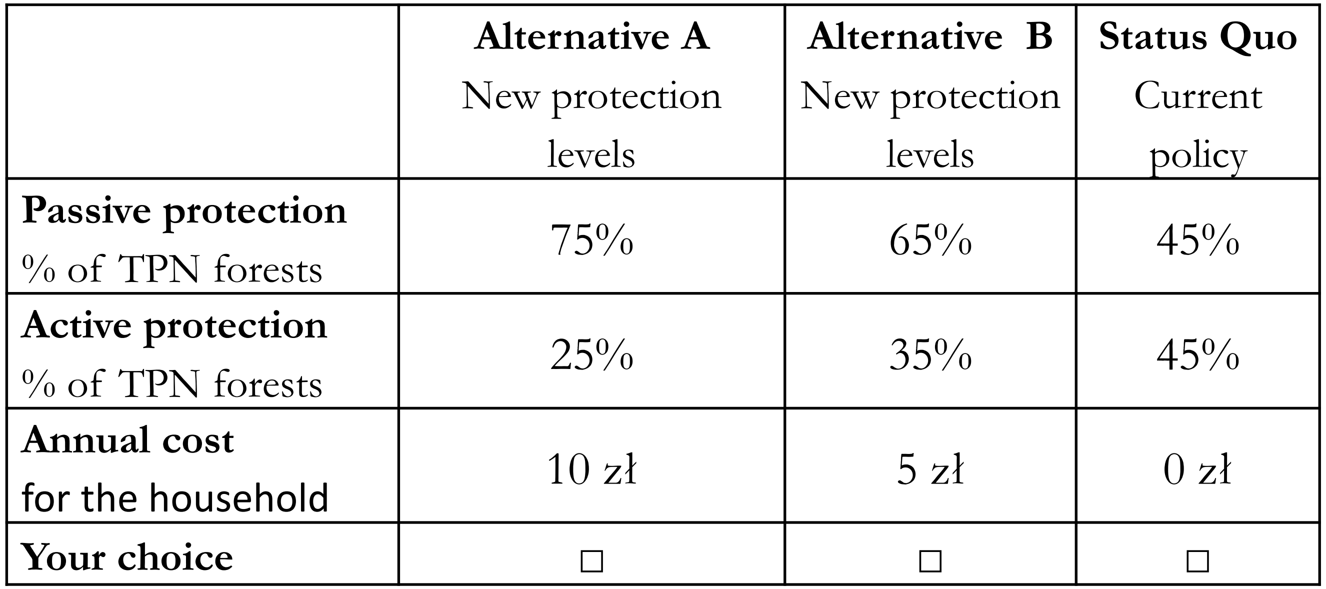
- It can be demonstrated that $\Sigma = \Gamma \cdot \Gamma'$ is a covariance matrix, so that $\boldsymbol{\beta}_i^r \sim MVN\left(\boldsymbol{\mu}, \Sigma\right)$

RANDOM PARAMETER MODELS

Controlling for correlations is important and often significantly improves model fit

- Correlation of tastes
- Scale heterogeneity / heteroskedasticity
- Single attribute being coded as multiple variables

EXERCISE 4: PASSIVE PROTECTION OF TNP



	Alternative A New protection levels	**Alternative B** New protection levels	**Status Quo** Current policy
Passive protection % of TPN forests	75%	65%	45%
Active protection % of TPN forests	25%	35%	45%
Annual cost for the household	10 zł	5 zł	0 zł
Your choice	□	□	□

1. Read *TPN.Rdata* into R
   - Stated preference study on passive protection in Tatra National Park
2. Estimate an error component model
3. Estimate a random parameter model with all coefficients being random
   1. Estimate a version with log-normal distribution for cost
   2. Estimate a version with correlated random parameters
4. Calculate median willingness to pay