DATA MINING

Regression with panel data

1 TOOLS

STATA® release 15

1.1 BACKGROUND

- Linear Regression:
  - For every linear model: $Y_i = \alpha + \beta * X_i + \mathbf{U}_I$
  - Goal: estimate a model that best fits the true model $\hat{\alpha}, \hat{\beta}$

- Methodology: OLS(Ordinary Least Squares)
  - To minimize the sum of the squares residuals:
    - $min \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^{2}$
    - Where $\widehat{Y_i} = \hat{\alpha} + \hat{\beta} * X_i$

2 METHODOLOGY

- 2.1 Linear Regression
- 2.1 Non-linear Regression

2.1 LINEAR REGRESSION

- The Fixed Effects Model
  - $\text{GDP}_{i,\ t} = \alpha + \beta_i * \text{X}_{i,\ t} + \Theta_t + \mathbf{U}_{i,\ t}$

- The Radom Effects Model
  - $\text{GDP}_{i,\ t} = \alpha + \beta_i * \text{X}_{i,\ t} + y_t * \text{E}_t + \mathbf{U}_{i,\ t}$

➢Where X is repressor, i is county, t is year, Θ is the fixed effects over years, E is dummy variable of the year.

FIXED VS RANDOM

	Pro	Con
Fixed	Can only see the time effect within-year	No assumption needs
Random	Efficient Clearly see the time effect between-year and within-year	We need to assume there is no correlation between time effect and regressor

FIXED OR RANDOM?

- Hausman test
  - H_0:no correlation between regressor and time effect

    or $\mathrm{cov}(X_i , X_{i,t}) = 0$

  - Under H_0: Random effects model is consistent and efficient, while fixed effects model is consistent but not efficient

  - Reject H_0: Random effects model is not consistent, but fixed effects model is still consistent

2.2 NON-LINEAR REGRESSION

- $\text{Log}(\ \text{GDP}_{i,\ t}\ ) = \alpha + \beta_i * \text{Log}(\ X_{i,\ t})$

3 RESULT

- Coefficient of determination:
  - Linear regression: 98.68%
  - Non-linear regression: 96.74%

- Final model:
  - Linear regression with panel data using fixed effect

GDP$_{i, t}$ = 1.85*10^11 – 1.8*10^11 * CO2 emission + 2.39 * foreign investment + 238.13 * labor force
+ 3.22*10^7 * technical articles