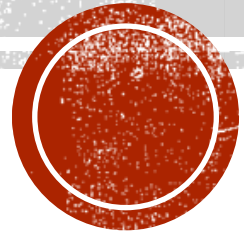
# DATA MINING

Regression with panel data



# 1 TOOLS

# 2 METHODOLOGY

- 2.1 Linear Regression
- 2.1 Non-linear Regression

# 2.1 LINEAR REGRESSION

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

- The Radom Effects Model
  - $GDP_{i,t} = \alpha + \beta_i * X_{i,t} + \gamma_t * E_t + \mathbf{U}_{i,t}$

➢ 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<br>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 $\text{cov}(X_i , X_{i,t}) = 0$

  - 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

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

# 3 RESULT

```
Fixed-effects (within) regression               Number of obs      =     1,562
Group variable: year                            Number of groups   =        11

R-sq:                                           Obs per group:
     within  = 0.9868                                         min =       142
     between = 0.9584                                         avg =     142.0
     overall = 0.9860                                         max =       142

                                                F(4,1547)          =  28818.17
corr(u_i, Xb)  = -0.0517                        Prob > F           =    0.0000
```

| gdpconstant2010us | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| co2emissionskgper2010usofgdp | -1.80e+11 | 5.79e+10 | -3.11 | 0.002 | -2.94e+11 | -6.68e+10 |
| foreigndirectinvestmentnetinflow | 2.386636 | .2353181 | 10.14 | 0.000 | 1.925059 | 2.848212 |
| laborforce | 238.127 | 84.36844 | 2.82 | 0.005 | 72.63836 | 403.6155 |
| technicalarticles | 3.22e+07 | 300933 | 106.94 | 0.000 | 3.16e+07 | 3.28e+07 |
| _cons | 1.85e+11 | 3.89e+10 | 4.76 | 0.000 | 1.09e+11 | 2.61e+11 |
| sigma_u | 2.397e+11 | | | | | |
| sigma_e | 9.781e+11 | | | | | |
| rho | .05668021 | (fraction of variance due to u_i) | | | | |

F test that all u_i=0: F(10, 1547) = 8.46 Prob > F = 0.0000

- Coefficient of determination:
  - Linear regression: 98.68%
  - Non-linear regression: 96.74%
- Final model:
  - Linear regression with panel data using fixed effect

GDP = 1.85*10^11 – 1.8*10^11 * CO2 emission + 2.39 * foreign investment + 238.13 * labor force + 3.22*10^7 * technical articles