## Day 2: Empiracal Analysis

**Agenda:**

Learn how to create do-files

Learn how to draw line plots between variables

Learn how to run benchmark regression

Learn how to run group regression

Learn how to perform robustness checks for panel data

1. Create do-files and record your commands

Way 1: click the button “New do-file editor”

Way 2: select the command from the history and send the command to the do-file editor.

1. Line plot

Line plot allows you to review the relationship between two variables based on each individual

----------------------------------------------------

xtline invest mvalue

------------------------------------------------

reghdfe areg

1. Group regression

Group regression allows you to run regression in different groups.

1. Use the command “merge” to include regional data
2. Transfer the string variables into numeric variables (egen g\_region = group( region ) )
3. label define:

----------------------------------------------------

label define lg\_region 1 "Asia" 2 "Europe" 3 "North America"

----------------------------------------------------

1. combine the label with variables:

----------------------------------------------------

label values g\_region lg\_region

----------------------------------------------------

1. drop original variable “region”

----------------------------------------------------

drop region

----------------------------------------------------

1. start group regression

----------------------------------------------------

bysort g\_region: xtreg invest mvalue kstock i.year,fe

----------------------------------------------------

1. if you want to output the results, use the command outreg2 to help:

----------------------------------------------------

bysort g\_region: outreg2 using grouping\_result.doc, replace pvalue bdec(3) pdec(3) keep(invest mvalue kstock) addtext(Time fixed effect, YES, Individual fixed effect, YES): xtreg invest mvalue kstock i.year,fe

1. Robustness check

To check whether your results are stable, you should: replace the dependant variables and core explanatory variables respectively; and solve the endogenous problems.

What is endogenous problems?

It originates from the sample selection bias, omitted variables and simultaneous causality.

How to solve endogenous problems?

We usually use instrument variables (IV) to solve endogenous problems. Other methods such as DID(Difference in differences), RDD(regression discontinuity design), PSM(propensity score matching) are also applicable to solve endogenous problems.

2SLS and GMM

1. 2SLS: principal and command

Example: graduation rate and education quality

----------------------------------------------------

webuse mroz,clear

global convars "exper expersq"

ivregress 2sls lwage $convars (educ = motheduc fatheduc),first

ivreg2 lwage $convars (educ = motheduc fatheduc),first

----------------------------------------------------

1. System GMM(Generalized Method of Moment)

----------------------------------------------------

webuse grunfeld,clear

sort company year

xtabond2 invest l.invest kstock , gmm( invest , lag(1 4) collapse) gmm( kstock , lag(1 4) collapse) iv( mvalue ) robust two

----------------------------------------------------

AR（1）<0.1

AR（2）>0.1

Hansen p-value: 0.1-0.25

## Heterogeneous analysis and mechanism test

Learn how to explain the results

Learn how to conduct heterogeneous analysis

Learn how to conduct mechanism test

1. **Heterogeneous analysis**

To explore whether the effect of X on Y will be stronger when a factor participates.

Y = β₀ + β₁X₁ + β₂X₂ + β₃ X₁\*X₂ + ε

Example: diligence, score and ability/interest

----------------------------------------------------

webuse grunfeld,clear

gen interact= mvalue\* kstock

xtreg invest mvalue kstock interact, fe

Quiz: run heterogeneous analysis based on development level

----------------------------------------------------

merge m:m company using xxx

drop \_merge

way1:

gen dev= (development == "developed")

label define l\_dev 1 "developed" 0 "developing"

label values dev l\_dev

way2:

egen g\_dev = group( development )

gen dev = 1

replace dev = 0 if g\_dev == 2

label define l\_dev 1 "developed" 0 "developing"

label values dev l\_dev

----------------------------------------------------

try to run group regression based on development

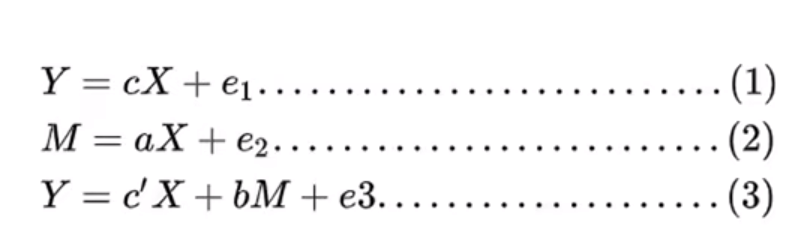
What is the difference between group regression and heterogeneous analysis?

1. **Mechanism test**

What is the mediating effect?

M

X Y



The coefficient **c** should be significant

The coefficient **a** and **b** should be significant

"complete mediation effect" and "partial mediation effect"

**Stepwise Regression:**

----------------------------------------------------

use "http://www.stata-press.com/data/r15/gsem\_multmed", clear

reg perform support

reg satis support

reg perform satis support

----------------------------------------------------

**Sobel—goodman mediation test + bootstrap:**

----------------------------------------------------

net install sgmediation2, from ("https://tdmize.github.io/data/sgmediation2")

Download ado file “sgmediation2” and paste into your folder “stata\ado\base”

sgmediation2 perform, iv( support ) mv( satis )

add your control variables into the cv(): sgmediation2 perform, iv( support ) mv( satis ) cv(\_id\* \_year\*)

bootstrap r(ind\_eff) r(dir\_eff), reps(500) : sgmediation2 perform, iv( support ) mv( satis )

## Other models

1. Probit and logit regression

Probit regression and logit regression are both methods used for modeling binary outcome variables, where the response variable can take on only two possible outcomes (usually coded as 0 and 1).

The Probit model assumes that the random disturbance term of the model follows a normal distribution; The random perturbation term corresponding to the Logit model is assumed to follow a Logistic distribution.

OLS→FE→RE

----------------------------------------------------

webuse union, clear

logit union age grade

probit union age grade

xtset idcode year

xtlogit union age grade,fe

xtprobit union age grade,fe

xtlogit union age grade,re

xtprobit union age grade,re

----------------------------------------------------

1. Tobit model

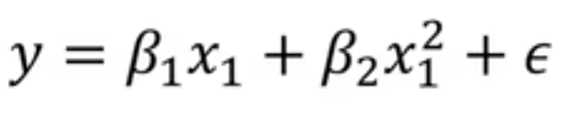
The Tobit model, also known as a censored regression model, is a statistical model used for analyzing data where the dependent variable is either left-censored, right-censored, or both.

----------------------------------------------------

sysuse nlsw88,clear

tobit wage hours ttl\_exp tenure , ll(0) ul(#) vce(cluster idcode )

1. U-shape model



----------------------------------------------------

ssc install utest, replace

clear

open the file “U-shape.dta”

twoway(scatter y x, mcolor(yellow)) (qfit y x)

applot y x,quadratic

gen x2=x^2

reg y x x2

utest x x2

sum x

----------------------------------------------------

If the "extreme point" detected is not within the data range, we cannot reject the original hypothesis.

To test whether there is a U-shape relationship between wage and ttl\_exp

----------------------------------------------------

sysuse nlsw88,clear

---------------------------------------------------

1. Loop

sysuse auto,clear

----------------------------------------------------

gen lnprice=ln(price)

generate lnmpg = ln(mpg)

generate lnrep78 = ln(rep78)

generate lnheadroom = ln(headroom)

generate lntrunk = ln(trunk)

generate lnweight = ln(weight)

generate lnlength = ln(length)

foreach x of varlist price mpg rep78 headroom trunk weight length{

gen ln`x' = ln(`x')

}

forvalue i=1(1)20{

gen x\_`i' =0

}

forvalue i=1(2)20{

drop x\_`i'

}

drop x\_\*