Lecture 5: Data Analysis

Yi Chen

ShanghaiTech University

2021

Outline

Outline

The # 1 principle in Mankiw's Ten Principles of Economics—People face trade-offs.

So do economists themselves.

- Which data to choose? (Lecture 2)
- How to deal with missing values? (Lecture 4)
- Which specification to use?
- How to weigh the sample?
- What is the most appropriate standard error?

Stata syntax for regression:

regress *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, noconstant vce(*vcetype*)]

In this lecture, we will focus on:

1. Choice of *depvar* and *indepvars*
2. Choice of *weight*
3. Choice of vce(*vcetype*)

Which Specification to Use?

- The answer from undergraduate-level econometrics class: R^2, adjusted R^2, AIC, BIC.
- These are *ex post* criteria.
- In actual research, it is more appropriate that specifications are determined in an *ex ante* fashion.

Choice of Dependent Variable

- Dependent Variable: log or level?

	Level			Log		
	Before	After	Δ	Before	After	Δ
Treat	1000	2000	1000	6.908	7.601	0.693
Control	600	1200	600	6.397	7.090	0.693
DID			400			0

- *Ex Ante* knowledge in what form the shock would take place.
- *Ex Post* check the common trend assumption.
  - In the above example, if the common trend prior to the shock takes place in log, the DID in levels gives wrong results (and vice versa).

Choice of Control Variables

- Too few control variables—omitted variable bias
  - But keep in mind that OVB is not always problematic.
- More control variables always better?
  - e.g., if you are estimating the returns to education, should industry dummies be controlled for?
- Bad control (Most Harmless, Reading Material 5.1)
  - "Bad controls are variables that are themselves outcome variables in the notional experiment at hand."
  - "Good controls are variables that we can think of as having been fixed at the time the regressor of interest was determined."
- Have a clear idea about the distinction between *control variables* and *mechanisms*.
  - If not so clear... robustness check.

Functional Form: Parametric versus Non-parametric

- Non-parametric estimation is currently the central research area in econometric theory and is becoming increasingly popular in empirical analysis.
  - Machine-learning can also be viewed as one specific approach of non-parametric estimation.
- Non-parametric estimation is not a panacea.
  - It also requires assumptions for identification purpose, although generally less strict compared with parametric models.
  - The technique itself says nothing to causal inference. For example, Amazon hiring A.I. learns gender discrimination against women.
- Curse of dimensionality greatly restricts its practical usage.
  - That's why non-parametric approach has been mainly adopted in regression discontinuity design.
- Even in non-parametric model, econometricians also face trade-offs.
  - e.g., optimal bandwidth, consistency versus accuracy.

Functional Form: Linear versus Non-linear

- The # 1 principle of correctly understanding non-linear model: non-linear model is not just fancy linear model.
- Conclusions derived from linear model, in many cases, cannot be directly applied to non-linear model.
  1. Normality assumption of the error term (misspecification).
  2. Forbidden regression (IV estimation with non-linear first stage).
  3. Probit model with fixed effects. Type -help xtreg- and -help xtprobit-, what difference do you find?
  4. Interpreting the interaction terms.

Interaction Terms in Probit Model

$$
\begin{aligned}
E(y|\mathbf{x}) &= \Phi(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2) = \Phi(.) \\
\frac{\partial^2 \Phi(.)}{\partial x_1 \partial x_2} &= \beta_{12}\Phi'(.) + (\beta_1 + \beta_{12} x_2)(\beta_2 + \beta_{12} x_1)\Phi''(.)
\end{aligned}
$$

Even if $\beta_{12} = 0$

$$
\frac{\partial^2 \Phi(.)}{\partial x_1 \partial x_2} = \beta_1 \beta_2 \Phi''(.)
$$

There is no guarantee that this values equals to zero. The function actually changes according to (x_1, x_2).

Solution:

- Average marginal effect (be careful with the interpretation)
- Use linear probability model

Personal recommendations to most non-technical practitioners

- Simply use the linear model if there is no clear drawback.
  - e.g., for binary endogenous variable, simply use linear probability model.
- In most cases, we have more important stuffs to worry about compared with the functional specifications, such as endogeneity, validity of common trend assumption.

Outline

Stratified Sampling and Sample Weight

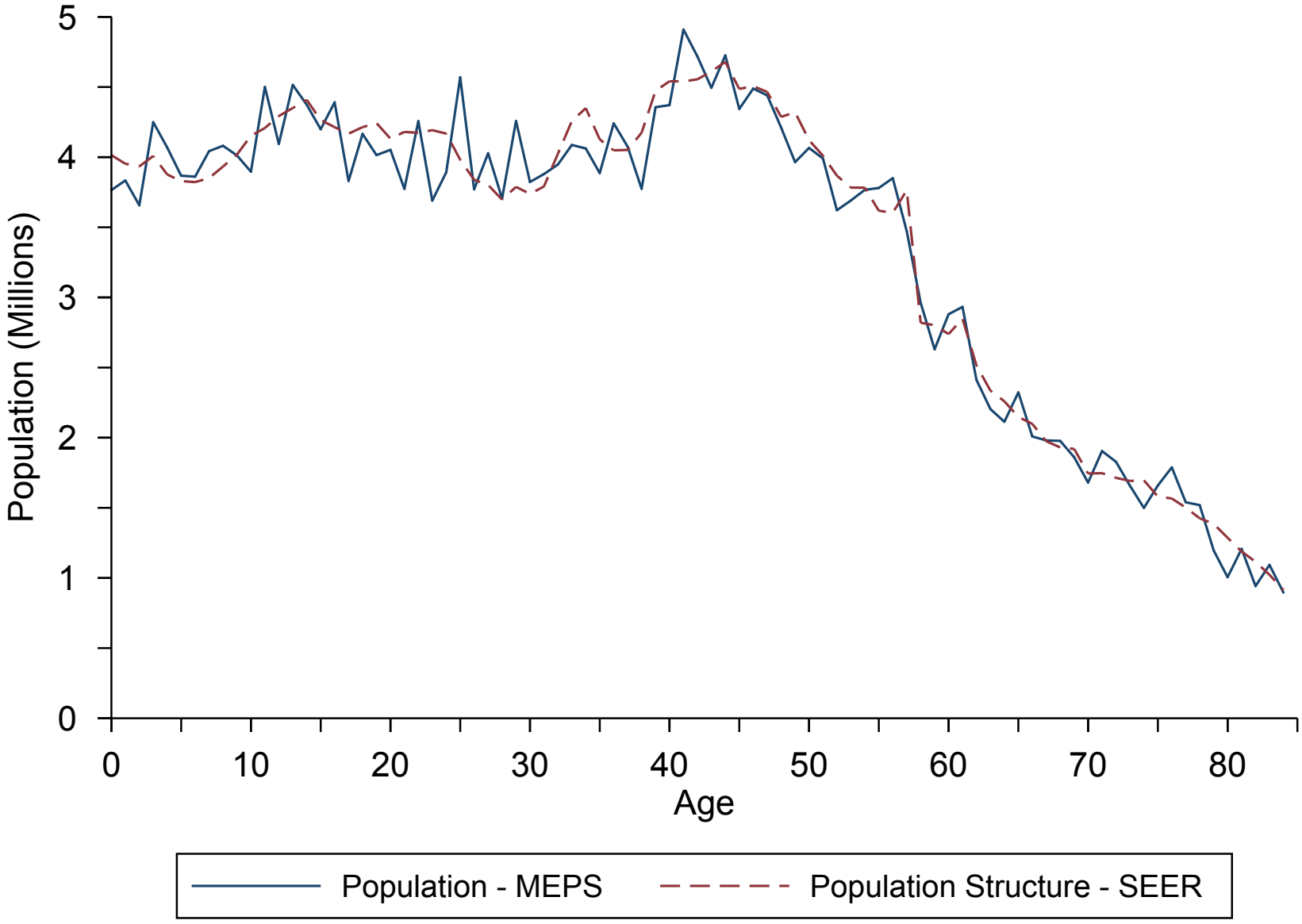
regress *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, noconstant vce(*vcetype*)]

- Stratification—the process of dividing members of the population into homogeneous subgroups before sampling.
  - The strata should be **mutually exclusive** and **collectively exhaustive**.
- Why stratification?
  - Some under-representative population may be of special interest, e.g., minority, income below poverty line.
  - The observations maybe not enough for convincing statistical analysis if using random sampling instead.
- Why weighing?
  - Back up the population average distorted by the oversampling of certain stratas.

Weight in Stata

- Two types of weight are often used in Stata. Frequency weight (*fweights*) and sampling weight (*pweights*). The usage of *fweights* is obvious and we will focus on *pweights*.
- While most commands for regression analysis support all types of weight, many commands for descriptive purpose do not support all types of weight.
  - -summarize- does not support *pweights*
  - -twoway graphs- does not support using weight.
- A universal solution: use -collapse- with *pweights* to construct the statistics first.

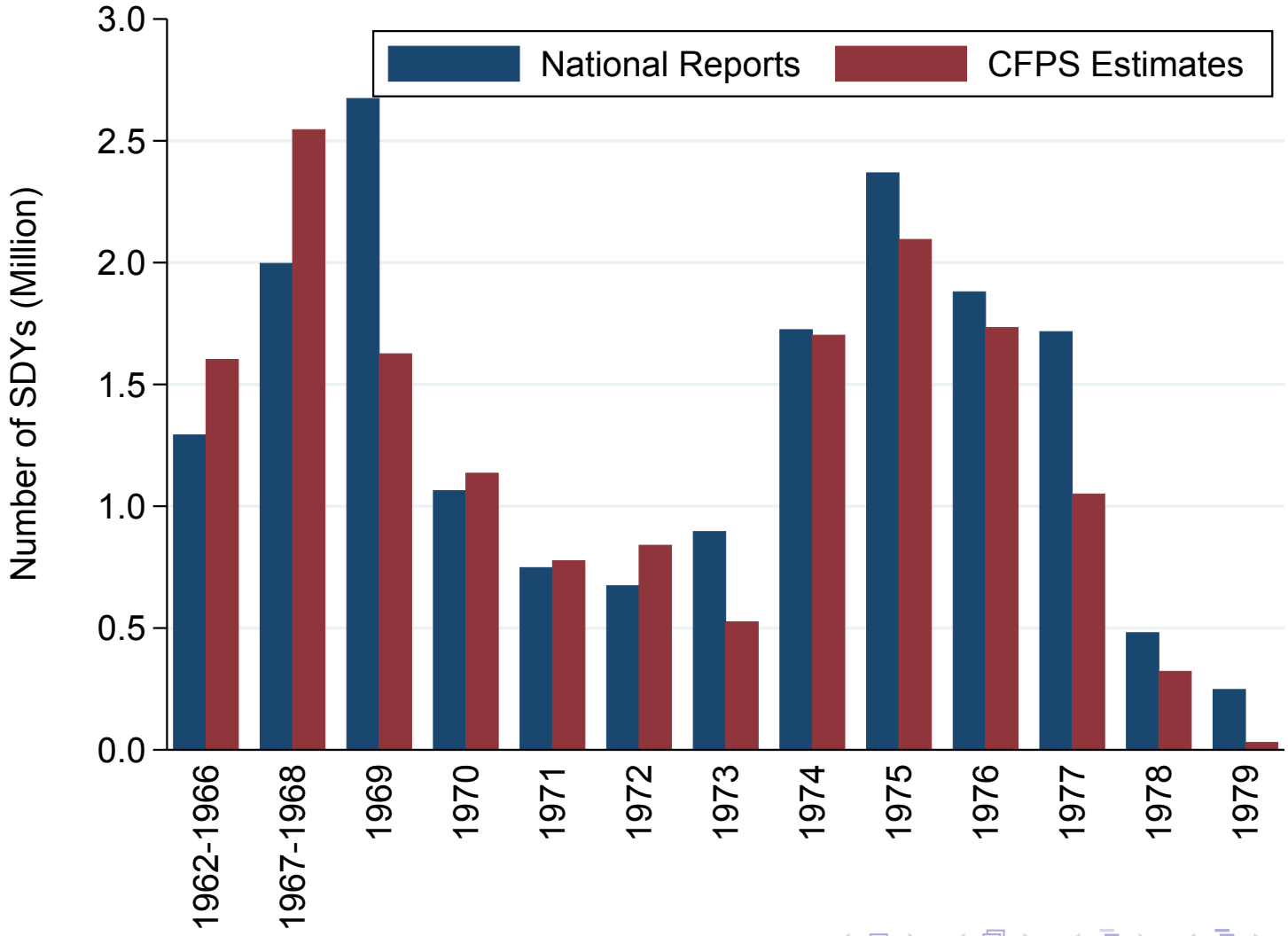
- *pweights*—"the inverse of the probability that the observation is included because of the sampling design."
  - *pweights* is usually a large number. It means how many national population this observation can represent according to the sampling process.
  - Stata example
- Descriptive analysis—effectively the same as frequency weight.
- Regression analysis
  - -reg- with sampling weight = weighted least square (use the square root of sampling weight as the weight) + robust standard error
- Stata example


Population (Millions)
0
1
2
3
4
5
0
10
20
30
40
50
60
70
80
Age
Population - MEPS
Population Structure - SEER

Application—Generating National Level Estimation

If you correctly understand the meaning of sampling weight, it can be used for applications beyond adjusting oversampling.

- e.g., back up TOTAL national counts from household survey.

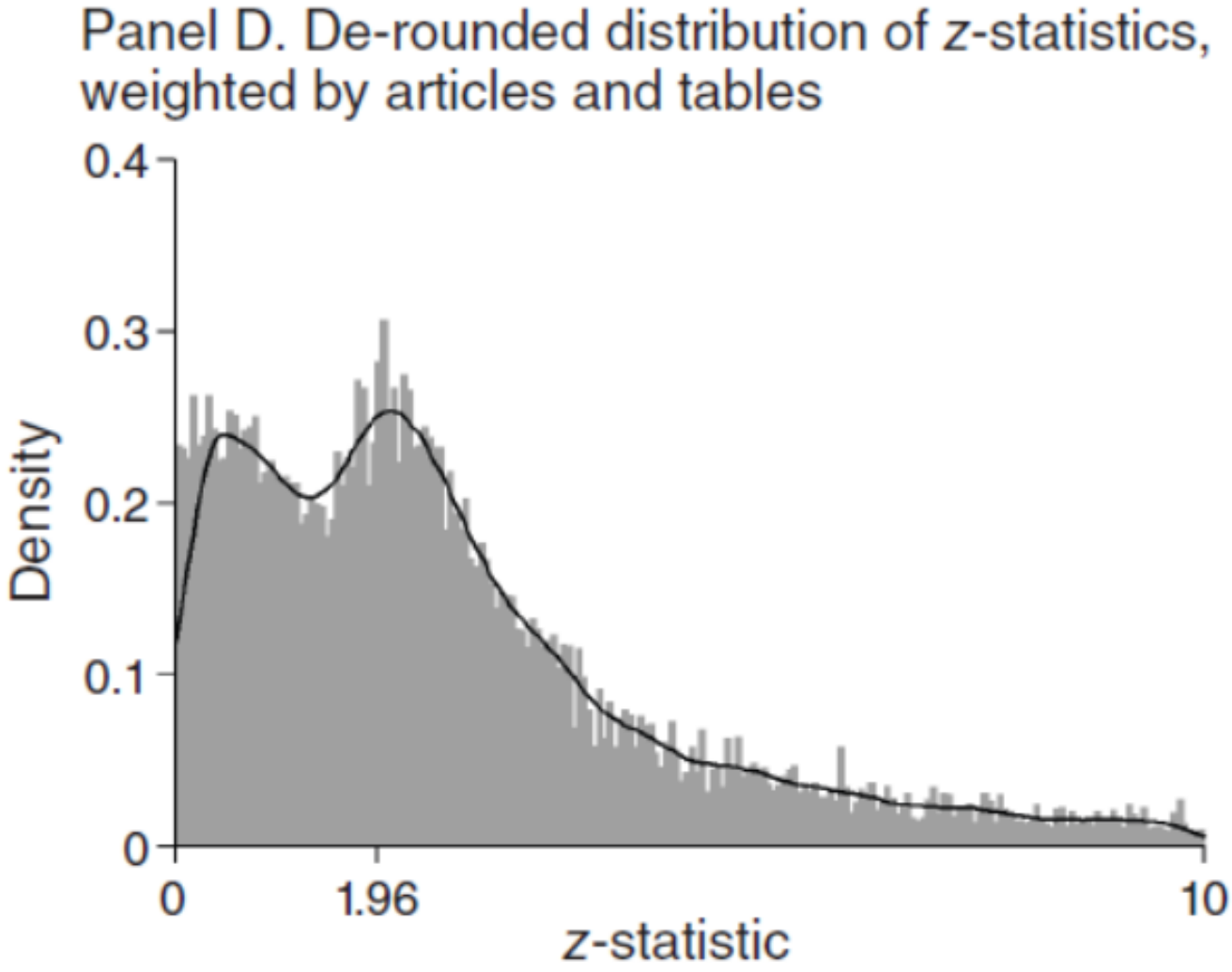


Is Sample Weighing Always Better?

- A simple answer–No.
  - Solon et al. (2015): "Be clear about the reason that you are considering weighted estimation, think carefully about whether the reason really applies, and double-check with appropriate diagnostics."
- If the data is homoscedastic, using the WLS results in efficiency loss. The loss could be larger than you expected.
- See the CFPS example.
  - There is no reason to believe the variance of the error term in a wage equation in Shanghai is ten times of that in Tianjin. Stata
- Read the user guide, understand the sampling framework, and think carefully whether you need to use the sampling weight!
- Reading Material 5.3 "What are we weighing for" for more details.

Standard Error and Clustering

- In econometric classes, we mostly focus on getting the estimated coefficients correct. Much less attention is given to estimating standard error.
- How important is standard error?
  - Brodeur et al. (2016) "Star Wars"
    - There is an increasing tendency to not reporting the stars (e.g., AER).
  - Bertrand et al. (2004): How much should we trust DID? Find an "effect" significant at the 5 percent level for up to 45 percent of the placebo interventions.
- Clustering at the proper level is now of crucial importance in the applied micro field.
- Reading Material 5.4 for more details.


Panel D. De-rounded distribution of z-statistics, weighted by articles and tables
0.4
0.3
0.2
0.1
0
Density
0
1.96
10
z-statistic

What is Cluster?

$$\Omega = diag\left(\Sigma_g\right)$$

$$= \begin{bmatrix}
\sigma^2_{(11)1} & \cdots & \sigma_{(1N_1)1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & & \vdots \\
\sigma^2_{(N_1 1)1} & & \sigma_{(N_1 N_1)1} & 0 & & 0 & & 0 & & 0 \\
0 & \cdots & 0 & \sigma^2_{(11)2} & \cdots & \sigma_{(1N_2)2} & & \vdots & & \vdots \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & & & \\
0 & \cdots & 0 & \sigma_{(N_2 1)2} & \cdots & \sigma^2_{(N_2 N_2)2} & & & & \\
 & & & & & & \ddots & \vdots & & \vdots \\
 & & & & & & & \sigma^2_{(11)g} & \cdots & \sigma_{(1N_g)g} \\
 & & & & & & & \vdots & \ddots & \vdots \\
 & & & & & & & \sigma_{(N_g 1)g} & \cdots & \sigma^2_{(N_g N_g)g}
\end{bmatrix}$$

When to Cluster?

- If you have aggregate variables
  - Especially if such variables are constructed or proxied (imagine what would happen if aggregate variables are measured with errors).
  - The role of cluster is different from that of dummy controls.
- Group-level shocks
- Serial correlation
  - For example, panel data with long T
- Similar to sampling weight, unnecessary usage of clustering leads to efficiency loss.
  - But empirically, there is little to lose as long as the number of clusters is large enough.

Cluster at What Level?

- Admittedly, often not that obvious...
- At the level at which you have aggregate variables
- The data might correlate in more than one way
  - If nested (e.g., household and county), cluster at the highest level of aggregation.
  - If not nested (e.g., time and space),
    - Include fixed-effects in one dimension and cluster in the other one.
    - Multi-way clustering extension (see Cameron et al. (2006))
    - Here it means cluster at time + space level, NOT time×space level!
- Official Stata command (such as -regress-) only allows one cluster variables.
- External command -reghdfe- supports multi-dimensional clustering.

Numbers of clusters & size of SE?

- Generally speaking, clustering would increase the estimated s.e..
  - Might decrease if the error terms are negatively correlated within a cluster.
- Generally speaking, the higher the clustering level, the larger the resulting s.e..
  - However... this is conditional on the s.e. is consistently estimated.
- The consistency relies on the number of clusters $g \to \infty$ as the sample size goes to infinite.
- Rule of thumb, $g \geq 50$
  - How to quickly know the number of clusters? External command -unique-
  - If g is small, use "wild bootstrap" (Cameron et al., 2008)

Outline

Mis-used Empirical Strategies

- There are two strategies that are considered as "cheap tools" to address endogeneity issue:
  - Arellano–Bond estimator (dynamic panel data estimator)
  - Propensity score matching (PSM)
- Those two strategies are very popular in Chinese publications, but 99% of papers are not using them correctly.
- In short, Arellano–Bond estimator is closer to fixed-effects approach; PSM is closer to regression.

PSM—OLS or IV?

- Let's see how Angrist & Pischke organize their *Most Harmless*.
- Matching helps control only for OBSERVABLE differences, not unobservable differences.
- PSM can be understood as a semi-parametric form of OLS.
- It is better to combine PSM with other techniques (to gain efficiency), e.g. PSM-DID
  - Note it is the DID part that gives you identification.

1. Matching的本质

国内很多文章都说matching能解决内生性问题，一直没有查到该说法的最原始出处，但是被国人以讹传讹流传到现在，所以在这里必须强调下matching和OLS一样都属于selection on observables的方法，识别假设都是是unconfoudness或者说是conditional independent assumption，（y0,y1），此外matching还多了个overlap（共同区间假设）。所谓CIA假设，就是控制住所有可观测因素x后，未观测变量不会对两组观测结果y产生系统性差异，也就是内生性问题。因为内生性问题意味着未观测因素是可能的confounding factor，既影响个体选择D，又影响结果变量y，即便控制可观测因素，选择性偏差仍然存在，因此我们说matching本质不能解决内生性问题。

See Reading Material 5.5 for details.

Arellano–Bond Estimator (Dynamic Panel Data Estimator)

Dynamic Panel Data (with one or more lagged dependent variables)

$$\begin{aligned} y_{i,t} &= \mathbf{X}_{it}\beta + \rho y_{i,t-1} + \alpha_i + u_{i,t} \\ \Delta y_{i,t} &= \Delta\mathbf{X}_{it}\beta + \rho \Delta y_{i,t-1} + \Delta u_{i,t} \end{aligned}$$

$E\left(\Delta y_{i,t-1}\Delta u_{i,t}\right) \neq 0$ by construction, why?

Idea: $E\left(\Delta y_{i,t-j}\Delta u_{i,t}\right) = 0$ for $j \geq 2$, use lagged y as instrument for $\Delta y_{i,t-1}$.

- Arellano–Bond Estimator solves the "endogeneity" of $\Delta y_{i,t-1}$, which is created by the dynamic panel data structure.
- It has nothing to do with the endogeneity of $\Delta\mathbf{X}_{it}$! Actually, the whole model rests on the exogeneity of $\mathbf{X}$.

- Don't believe any universal solution to the endogeneity problem.
- The solution has to depend on the question/background/design of research.
- That's why we need RCT/DID/IV/RD...

EKONOMIPRISET 2019
THE PRIZE IN ECONOMIC SCIENCES 2019

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Abhijit Banerjee

Esther Duflo

Michael Kremer

"för deras experimentella ansats för att mildra global fattigdom"

"for their experimental approach to alleviating global poverty"

A bit More on IV Estimates

- First stage
- Endogeneity test
- Overidentification test
- Nonlinear IV

First-stage

- ALWAYS report the first stage.
- Rule of thumb that the F-statistics should be ≥ 10 is widely known.
- Now it is more common to adopt Stock and Yogo (2005), whose criteria is generally stronger and adjusts to the number of instruments.
- The external command -ivreg2- includes the critical values from Stock and Yogo (2005).

Endogeneity Test

- In undergraduate econometrics, you might have heard so-called "endogeneity test" named after a set of people (e.g., Durbin-Wu-Hausman test).
- But when reading recent literature, do you see modern papers do such kind of tests? No! Why?
  - Because they are useless!
- Essence of "endogeneity test"—assume IV is valid and tests whether the IV estimates statistically differs from OLS estimates.
  - However, the #1 concern is the validity of the instruments!

Overidentification Test

More useful than endogeneity test, but not as useful as many researchers' belief.

- Essence of "overidentification test":
  - If the number of instruments exceeds the number of endogenous regressors, you can identify the coefficient using different combinations of the instruments.
- "Overidentification test" says different combinations gives statistically indistinguishable results.

- Many people take the belief that it lends support to the credibility of instruments—if not pass, then at least one instrument is problematic.
- The rejection may also stem from that different sets of instruments identify different LATE.
- Good IV if pass?—in most cases, the extra instruments are constructed following a similar spirit. See the Stata example.
  - Could be very useful if different instruments are constructed from different perspectives.
  - But most researchers get exhausted with just one instrument...

Nonlinear IV

Again, non-linear model is not just fancy linear model! Even seemingly trivial model could be complicated in a non-linear set-up.

- Athey and Imbens (2006): *Econometrica*, "Identification and Inference in Nonlinear Difference-in-Differences Models"

Therefore, do realize nonlinear IV can be surprisingly difficult in techniques.

- Even more difficult in weak instrument test, overidentification test...
- More complicated standard error, such as clustering.

Example: Binary Regressor and Regressand

- continuous regressand + continuous regressor—standard IV
- binary regressand + continuous regressor—official Stata command -ivprobit-
- continuous regressand + binary regressor
  - Forbidden regression: run a first-stage Probit, get $\widehat{D}$ and plug into the second-stage.
  - Angrist and Pischke (2009) recommends that instead of directly plugging $\widehat{D}$, use $\widehat{D}$ as the instrument for the endogenous D.
- binary regressand + binary regressor — bivariate probit (Wooldridge (2010) Chapter 15 Binary Response Model)
  - Stata command: biprobit (Y = X R) (R = X Z)
  - Requires strong assumption about the distribution of the error terms (joint normality)

A Brief Summary of IV Estimation

- Critical assumptions: relevance, exogeneity, exclusiveness
- Relevance—can be explicitly tested, therefore, always report the first-stage.
- Exogeneity—can be partially tested, depending on the type of the instrument.
  - Overidentification test.
  - Randomized trial: balance check.
  - DID: common trend.
  - RD: no discontinuity in other covariates.
- Exclusiveness—very hard to test explicitly, especially if the channeling variables are not observed.

Lists of Non-linear Regressions in Stata

- -tobit- (linear regression with censoring data)
- -qreg- (quantile regression)
- -probit-/-logit-
- -oprobit-/-ologit- (ordered y variable)
- -mlogit- (y variable w/o natural ordering)
- -poisson- (count variable)
- -stcox- (Cox proportional hazards model for survival analysis)

Several Powerful External Commands

- -reghdfe-: "a generalization of areg (and -xtreg,fe-, -xtivreg,fe-) for multiple levels of fixed effects (including heterogeneous slopes), alternative estimators (2sls, gmm2s, liml), and additional robust standard errors (multi-way clustering, HAC standard errors, etc)."
- -ivreg2-
  - If you are just interested in the regression itself, -reghdfe- is enough. -ivreg2- is more powerful in the postestimation, such as the overidentification test.
- -rdrobust-: local polynomial regression-discontinuity estimation
- -psmatch2-: propensity score matching

Running Regression "Efficiently"

I am not going to talk about the technical details on regression...

Stata help file/Stata Journal/your econometric instructor should do a better job.

So...What I am going to talk about?

- Mastering interactions
- Mastering fixed effects
  - Fixed effects as control variables
  - Fixed effects as the variables of interest
- Use -predict- smartly

Mastering Interactions

- In many cases, we need to work with the interactions, such as DID & DDD.
- Of course, we can always generate the interactions by hand and then use it as a normal variable.
- Limitation
  - Tedious, especially if there are many categories.
  - Easy to make mistakes
    - Forget to exhaust all possible interactions.
    - In a placebo test, once you define a new *treat*, you also need to regenerate ALL the interaction terms.
- Actually, Stata provides convenient ways for interactions.

I borrow syntax for -reghdfe- *absvar*. The syntax works perfectly well as the control variable.

General rule: c. - continuous var; i. - indicator var; # interaction; ## exhaustive interaction

absvar	Description
i.*varname*	categorical variable to be absorbed (the i. prefix is tacit)
i.*var1*#i.*var2*	absorb the interactions of multiple categorical variables
i.*var1*#c.*var2*	absorb heterogeneous slopes, where *var2* has a different slope coef. depending on the category of *var1*
var1##c.*var2*	equivalent to "i.*var1* i.*var1*#c.*var2*", but *much* faster
var1##c.(*var2 var3*)	multiple heterogeneous slopes are allowed together. Alternative syntax: *var1*##(c.*var2* c.*var3*)
v1#*v2*#*v3*##c.(*v4 v5*)	factor operators can be combined

Mastering Fixed Effects I—FE as Control Variables

FE as control variables implies we are not interested in estimating the fixed effects.

- xi: regress y x i.FE—No!
  - It first generates a list of dummy variables and then put it in the regression. Very slow if you have large sample size or many categories.
- regress y x i.FE—if you are only interested in the coefficient of x, No!
  - It hinders you from quickly viewing the results if there are many categories.
- reghdfe y x, absorb(FE)—Yes!
  - Combine with the usage of interactions, you can easily realize more complicated controls of fixed effects, such as province×year.

Costs of Too Many Fixed Effects

In a linear model, if you are not interested in the fixed effects, the number of fixed effects is generally not an issue.

- Think about individual fixed effects in a panel model.
- The cost is fundamentally the same as adding extra variables
  - Loss of degree of freedom.
  - If the extra fixed effects do not capture additional variations in the residual term, it increases the standard error.

However, too many fixed effects will be an issue if...

- You are interested in the fixed effects.
- You are running non-linear models, such as Probit.

Fixed Effects in Non-linear Models

- If the FE can perfectly predict the outcome variable in a subcategory, you effectively lose all the observations in that category. Stata example
- If a non-linear Stata command does not converge, 99% cases are because there are too many fixed effects (or some variables predict the outcome variable too well).
- In what scenarios would this happen?
  - Only one observation in a subcategory.
  - No variation in the outcome variable within a subcategory.

Solution

- Control FEs at a higher level.
- Combine smaller categories into a large category.
- Drop the observations that have no variation in the outcome variable within a subcategory.
  - Q: how to do that?
  - A: -egen sd()-

Mastering Fixed Effects II—FE as the Variables of Interest

FEs do not merely serve as control variables. In many scenarios, FEs could be the variables of interest. For example,

- Life-cycle profiles (age FEs)
- Business cycle (year FEs)
- Spatial characteristics (city FEs)

Interpretation of FEs relies crucially on the "reference group." Only relative values matter in most cases.

Example: Gender Segregation in the Labor Market

Gender Segregation: women on average earn lower wages because they are more concentrated in low-paying occupations (or industry). But how to define low-paying occupations?

Occupation-average wages suffer from a reverse causality issue. An occupation is low-paying because there is more women.

A more scientific way:

$$\log\left(\text{Wage}\right)_i = \theta_0 + \theta_1 \text{Female}_i + \sum_d \theta_{2,d} \text{Occupation}_{i,d} + \varepsilon_i.$$

$\theta_{2,d}$ can be viewed as the occupation characteristics after controlling for gender composition.

If FEs are the variables of interest, you need to guarantee the FEs can be identified.

- Take a two-wave balanced panel data for example ($N \times 2$ observations).
- Although it is entirely feasible to estimate an individual fixed effect model, you cannot obtain the FEs.
- Therefore, you need to make sure the sample size in each cell is large enough.
  - Q: how to drop the cells with observations ≤ 50?
  - A: -egen count()-

How to Get the FEs?

- Method 1: regress y i.age
  - The first category will be automatically omitted because of the perfect collinearity.
  - Attention! _b[26.age] is not the FE for age 26!
    - _cons + _b[26.age] is.
- Method 2: reghdfe y, absorb(agefe = age)
  - This creates a new variable named "agefe", which equals to the age fixed effects of the corresponding observations.
  - The level of the "agefe" is adjusted so that the sample average is zero.
  - Don't forget to add back the constant term (latest version of -reghdfe- reports the constant term).
  - Stata example

Advanced Usage of -predict-

- Basic usage: -regress-, then -predict- within the same data set.
- Actually, -predict- can be viewed as a mapping from *varlist* + regression to predicted outcomes.
  - *varlist* and regression do not need to come from the same source.
- Application 1: Work with different data sets.
  - We wish to know the mortality rate in MEPS. But MEPS is not suitable for estimating the mortality.
  - Use NHIS, estimate the mortality rate as a function of health and age.
  - Store the estimation.
  - Open MEPS $\rightarrow$ load the estimation $\rightarrow$ make sure the variable names in MEPS are the same as NHIS $\rightarrow$ predict
  - See the Stata example.

Application 2: Blinder-Oaxaca Decomposition

$$
\begin{aligned}
\log W_m &= \alpha_m + \theta_m \mathbf{X}_m + \varepsilon_m \\
\log W_f &= \alpha_f + \theta_f \mathbf{X}_f + \varepsilon_f \\
&\downarrow \\
\Delta &= E\left[\log W_m - \log W_f\right] \\
&= \left(\alpha_m - \alpha_f\right) + \left(\theta_m E\left(\mathbf{X}_m\right) - \theta_f E\left(\mathbf{X}_f\right)\right) \\
&= \theta_m \left[E\left(\mathbf{X}_m\right) - E\left(\mathbf{X}_f\right)\right] + \left[\left(\alpha_m - \alpha_f\right) + E\left(\mathbf{X}_f\right)\left(\theta_m - \theta_f\right)\right]
\end{aligned}
$$

"Explained part" + "Unexplained part" (often interpreted as "discrimination" in early studies, be very careful with this interpretation.)

How do we get the second part from Stata?

B-O decomposition is now rarely used as the core part of a paper, but the idea is still highly relevant.

Outline

Simulation Method in Stata

- We are not always using data from real world.
- Simulated data is useful when we don't know the DGP (data-generating process):
  - Placebo test
  - When closed-form expression is difficult (e.g., bootstrap)

Placebo Tests

As can be seen from previous IV discussion, it is very difficult to defend the exogeneity assumptions of the instruments. Placebo tests are useful tools. Take DID for example,

- If you are worried about the common trend and time-invariant unobservable local characteristics.
- Try to shift the reform years back/forth by a couple of years to see if the effects are still there.

Placebo tests do not say about

- The exclusiveness restriction.
- Time-variant unobservables.

Aside from moving the reform years back and forth, assigning placebo treatment (shuffling) is also a common practice.

- See the Stata example for two types of placebo tests.
- Don't forget about the replicability of placebo assignment.
  - -set seed-
  - Store the placebo assignment in an external file

Even if placebo tests suggest that the random assignment does not generate your results, they do nothing about whether the estimated effects come from the treatment or some unobservables combined with the treatment.

A Quick Introduction to Bootstrap

- What is bootstrap? Lifting yourself.
- Datasets are often samples of the populations. Standard error captures the uncertainty from the sampling process.
  - Bootstrap views the sample you have in your data set as the population of interest.
- Why bootstrap?
  - Sometime it is difficult to get the analytical expression of the s.e.
  - The analytical expression often require certain assumptions. If those assumptions are violated, analytical expression may give wrong s.e. (cluster for example).
  - Bootstrap is believed to have better small sample performance.
- Another simulation method sharing a similar spirit: jackknife.

Standard Procedure

Original sample size N

1. Draw a sub-sample of size N WITH replacement from the original sample.
2. Calculate the desired statistics for the sample.
3. Repeat step 1 and step 2 for M times, where M is a large number.
   - Personal recommendation: 1,000 during preliminary investigation, 10,000 for the final output.
4. Treat one iteration as an observation, and calculate the s.e. in the normal fashion.

Miscellaneous Issues

- Bootstrap is only used to calculate the standard error. The point estimate is always the estimates from the original sample.
- Bootstrap method heavily depends on the independence assumptions across observations (or clusters).
  - When bootstrap for an individual panel data ($N \times T$), draw all T years for an individual, don't draw on individual-year.
  - If the data is clustered, draw the whole cluster all together!
- More technical issues
  - Bootstrap assumes the existence of the second moment.
  - Bootstrap assumes a speed of convergence at $N^{-\frac{1}{2}}$ (most of the estimators we learn so far converge at the speed).

Stata Implementation

There are three ways for Stata to perform bootstrapping method

1. Commands that incorporate the bootstrap option, such as -regress-
   - -vce(bootstrap)- or -vce(bootstrap, cluster(*varlist*))-
2. Commands plus the return list
   - bootstrap *exp_list* [, options eform_option]: *command*
3. Sometimes the statistics is not readily available in the return list, in that case, you may need to write your own program.
4. Stata practice.

Exercise: "How Much Should We Trust DID Estimates?"

Simulated Data: sim_wage_0.8 is generated in following steps:

1. Generate a state-by-year data with 1,050 observations/cells. There are 50 states and 21 years ranging from 1979 to 1999.
2. For each cell, generate a pseudo "education" that is uniformly distributed in (7,9).
3. The dependent variable (log_wage) is generated with the following equation:

$$\text{log\_wage}_{st} = 8 + 0.1 \times \text{Edu}_{st} + \varepsilon_{st}.$$

Subscript s and t represent state and year, respectively. ε_{it} is homoscedastic with a standard deviation of 0.1 and is independent across state (corr(ε_{it}, ε_{jp})=0 for any t and p if $i \neq j$). But it is serially correlated within state (corr(ε_{it}, $\varepsilon_{i,t-1}$)=0.8).

Serial Correlation and Overrejection

Generate a fake "law". First, draw a year at random from a uniform distribution between 1985 and 1995. Second, select exactly half the states (25) at random and designate them as "affected" by that law. People will be affected by the law ($I_{st} = 1$) if and only if they live in an affected state after the intervention date. Lastly, for each random draw, run the following regression:

$$\text{log\_wage}_{st} = \alpha + \beta \text{Edu}_{st} + \gamma I_{st} + \delta_s + \tau_t + \varepsilon_{it}.$$

δ_s and τ_t are the state fixed effects and year fixed effect, respectively. γ is the difference-in-difference estimator of interest.

Repeat the above simulation for 100 times and count how many times the estimator $\widehat{\gamma}$ is statistically significant at 5% if we ignore the serial correlation and assume i.i.d?

Reference

Angrist, J., Pischke, J.-S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion, 1st Edition. Princeton University Press.

Athey, S., Imbens, G. W., 2006. Identification and inference in nonlinear difference-in-differences models. Econometrica 74 (2), 431–497.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How Much Should We Trust Differences-In-Differences Estimates? The Quarterly Journal of Economics 119 (1), 249–275.

Brodeur, A., Lé, M., Sangnier, M., Zylberberg, Y., 2016. Star wars: The empirics strike back. American Economic Journal: Applied Economics 8 (1), 1–32.

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2006. Robust inference with multi-way clustering. Working Paper 327, National Bureau of Economic Research.

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2008. Bootstrap-based improvements for inference with clustered errors. The Review of Economics and Statistics 90 (3), 414–427.

Solon, G., Haider, S. J., Wooldridge, J. M., 2015. What are we weighting for? Journal of Human resources 50 (2), 301–316.

Stock, J. H., Yogo, M., 2005. Testing for Weak Instruments in Linear IV Regression. Cambridge University Press, pp. 80 – 108.

Wooldridge, J. M., December 2010. Econometric Analysis of Cross Section and Panel Data. No. 0262232588. The MIT Press.