Econometric Methods: Introduction

Maggie McConnell

Department of Global Health and Population

Harvard T. H. Chan School of Public Health

Plan of presentation

1. Intro to econometrics
2. Internal and External Validity
3. Applied examples
4. Some norms in Econometrics versus Biostatistics and Epidemiology
5. Learning to use a method + Continuous Development of Methods
6. Additional Slides
   - Epi-Econ Dictionary
   - Estimation methods: minimizing sum of squares, maximum likelihood, method of moments

Break for Questions

What is the point of econometrics?

- Econometrics is statistical methods used by economists
- The primary objective of econometrics is to identify *causal effects*
- What differentiates econometrics from other causal inference approaches is extremely limited tolerance for confounding
- A few important types of confounding
  - *Selection:* Who *choses* to participate in a program? Who responds to policy changes?
    - observable characteristics unlikely to fully explain human choices
  - *Reverse causality:* Does housing instability cause adverse health outcomes or do adverse health outcomes lead to housing instability?
    - Longitudinal data can help but causality often goes in both directions
    - People may foresee policy changes / environmental and change behavior before they occur

What is the point of econometrics?

- For questions of interest to economists, selection/”endogeneity” are fundamental
  - Underlying theoretical questions revolve around human behavior
  - Economics focused on making policy recommendations
    - Requires understanding direction of causality
    - Economics is less good at describing a problem

- Instead of trying to model or control for confounding + selection – econometrics develops techniques to avoid

Econometrics: focus on internal validity

- Ultimately economists require methods that ensure internal validity
  - Sometimes external validity is less of a focus
- Internal validity
  - The estimates are close to the true parameter values in the sample for which we have data
  - If methods are not internally valid, economists not interested in results
- External validity
  - The estimates are close to the parameter values in a population when they are used to predict outcomes
  - Generalizability and transportability
    - This really matters for thinking about scaling programs, understanding whether policy makers should use results from one setting/population to make more general policy

Markers of internally valid models

- Unbiased estimator
  - Expected value of estimated parameter equal true value in the population
- Consistent
  - An estimator that converges in probability to the population parameter as the sample size grows without bound
- Efficient
  - An estimator with the lowest variance relative to the population parameter among all possible estimators
- Correct confidence intervals
  - Needed for valid hypothesis testing
  - Often requires to thoughtfulness about what unit / cluster is treated by a program/intervention
- See notes for formal definitions

Internal validity

- Internal validity usually requires some assumptions
- However, only some of the assumptions are testable
  - Internal validity often defended by plausibility “norms” which differ by discipline
- Where a choice of methods is available, economists always prefer the method with fewer untestable assumptions
- However, extra assumptions may allow more efficient estimation if they are true
- New methods sometimes have “hidden” assumptions that are stated in highly technical language --may take time and examples to uncover plausibility

External validity

- Causal inference analyses provide estimates of causal effects for specific sample at a particular time under certain conditions
- Estimating causal impact of a specific intervention or policy – what about similar interventions/policies? differences in implementation?
  - Common pattern: positive results in early trials then limited impact when scaled up
- Would results apply to a broader population?
  - Random sample of population ensures parameters representative of the population
  - Special inclusion criteria for (especially common in RCTs) may limit external validity
    - i.e. drug trials on people with no co-morbidities, exclusion of some age groups / populations
- Some econometric techniques require very particular data/setting which may not always be consistent with externally valid samples
  - More discussion later
- Break for questions

Econometric strategies

- The objective of econometrics is to uncover causal impacts using methods with minimal assumptions
- One way to overcome problems of confounding – randomized controlled trial (RCTs)
  - With some assumptions comparing outcomes for treated and non-treated individuals allows us to uncover causal impacts
- Many econometric strategies we will cover in this segment try to “mimic” RCTs by creating a “counterfactual”
  - *Counterfactual:* what would outcomes have been in the absence of an intervention or policy change
    - Problem – we can’t observe the counterfactual
    - Methods we will cover in all represent different ways of constructing a counterfactual

Randomized trials: internal validity

- Randomized trials often referred to as gold standard – but they require assumptions for internal validity!

- Randomization asymptotically removes correlation between treatment and potential confounders
- Stable unit treatment value assumption
  - Potential outcomes for any unit don't vary with the assignment of other units
  - No spillovers – treatment of one unit does not affect other units, example of a spillover?
    - One solution is to change design – e.g., group randomization
- Successful randomization
  - Check balance tables
- Bias in reporting - was the trial double blind?
- Bias due to selective follow up – what was attrition rate? – was it differential between arms?
  - Very hard to fix analytically, much better to avoid

Application – evaluating the impact of interventions during pregnancy

- Research question: What is the impact of home-based support during pregnancy for low-income families?
  - Suppose that we care about is an outcome related to how much *prenatal care* pregnant people get
- We will:
  - Discuss a propensity matching approach
  - Look at results from an RCT on a related program in a similar population
  - Talk through how other causal inference approaches might be utilized

Propensity score matching example

- Program: Midlands Healthy Start (MHS)
  - Objective: increasing access to prenatal care; and removing barriers to health care access
  - Services: case management, outreach and recruitment, health education, a local health system action plan
- Evaluation strategy
  - Vital records data for all singleton births in counties where the program was operating
  - First step: compare births to families participating in MHS to those not participating
  - Next step: propensity score adjustment

Results

- Let’s compare the rate of “inadequate” prenatal care across participants and non-participants first

	Non-MHS rate (non-treated)	MHS rate (treated)
"Inadequate" prenatal care	17.0%	23.2%

- Why would program participants have higher rates of “inadequate” prenatal care?

Characteristic	Pooled Sample		
	Non-MHS (n = 48 826)	MHS (n = 7203)	Standardized Difference[a]
Mean maternal age, y	27.9	24.8	−0.55
Mean maternal prepregnancy BMI	27.5	28.7	0.15
Maternal age, y, %			0.56
< 20	6.5	17.2	
20–24	23.6	37.4	
25–29	30.8	25.6	
30–34	25.9	13.1	
≥ 35	13.2	6.7	
Maternal race/ethnicity, %			1.00
Non-Hispanic White	60.2	18.9	
Non-Hispanic African American	32.4	74.5	
Hispanic	4.2	5.7	
Non-Hispanic other	3.1	1.0	
Maternal education, %			0.65
< high school	9.6	22.4	
High school or equivalent	19.8	31.0	
Some college	36.7	36.8	
Bachelor's degree	21.0	7.0	
Graduate school	12.9	2.8	
Medicaid medical insurance, %	40.3	79.9	0.88
Nulliparous, %	41.7	49.2	0.15
Prepregnancy BMI category, %			0.20
Underweight (< 18.5 kg/m^2)	3.2	4.1	
Normal weight (18.5–24.9 kg/m^2)	39.1	33.6	
Overweight (25.0–29.9 kg/m^2)	24.9	24.5	
Obese (≥ 30.0 kg/m^2)	29.0	36.4	
Missing	3.9	1.4	
Smoking before pregnancy, %	12.4	15.8	0.10
Composite disease history score ≥ 1, %[b]	25.2	23.5	−0.04

Results

- Now let's use propensity score matching to compare only those with similar observable characteristics

	Propensity Score Adjusted Odds Ratio
"Inadequate" prenatal care	0.86 (0.80, 0.92)

- Conclusion: "the Healthy Start program contributed to significant improvements in prenatal care"
- What do you think are the challenges to causal inference here?
  - What if individuals who *choose* to participate in the program are different on characteristics we haven't measured?
- How big of a problem do we think unobservable characteristics are in this setting?

Randomized controlled trial

Effect of an Intensive Nurse Home Visiting Program on Adverse Birth Outcomes in a Medicaid-Eligible Population

A Randomized Clinical Trial

Margaret A. McConnell, PhD[1,2]; Slawa Rokicki, PhD[1,3]; Samuel Ayers, BA[4]; et al

» Author Affiliations | Article Information

JAMA. 2022;328(1):27-37. doi:10.1001/jama.2022.9703

- Program: Nurse Family Partnership
  - Objective: improving health outcomes during pregnancy including supporting access to health care services
  - Services: home visits where nurses provide case management, outreach and recruitment, health education

Evaluation strategy

- Randomized controlled trial assigned eligible families to treatment or control
- Vital records data from counties where the program was operating
- Compare control group outcomes to treatment group outcomes

Exhibit 1. Characteristics of the Analytical Sample at Baseline & Balance Across Treatment Groups (% of sample)

	Full Sample (N=4,587)	Control Group (N=1,522)	Treatment Group (N=3,065)
Age at trial enrollment			
15-18 years	18.6%	18.2%	18.8%
19-24 years	54.9%	55.5%	54.6%
25-34 years	24.2%	24.7%	23.9%
35-44 years	2.3%	1.6%	2.7%
Race/Ethnicity			
Black non-Hispanic	52.0%	51.9%	52.1%
White non-Hispanic	32.2%	31.6%	32.6%
Hispanic	5.5%	5.9%	5.3%
Asian, multiracial, or other	4.1%	3.5%	4.3%
Educational Attainment			
Less than high school diploma	22.9%	22.4%	23.2%
High school diploma or equivalent	35.9%	34.6%	36.5%
Some college, less than bachelor's degree	33.8%	35.6%	32.9%
Bachelor's degree or higher	6.9%	7.0%	6.9%

Results

- Let’s compare the treatment group and control group rates of “inadequate” prenatal care

	Control group	Treatment group
"Inadequate" prenatal care	18.70%	19.60%

- No statistically significant difference
- Previous study concluded that a similar program reduced likelihood of receiving ”inadequate” prenatal care
- What are some reasons the result is different?
  - External validity: important programmatic differences
  - What about “unobserved/unmeasured characteristics” correlated with both program participation and receiving antenatal care? Examples?

Results

- We can use this setting to learn something about *how* different outcomes are for program participants compared to non-participants
- Eligibility criteria: first-time birth, Medicaid eligible, >=15 years old and living in counties where program is operating
- We can observe the outcome among eligibles who did not participate

	Control group	Treatment group	Non-participant but eligible
"Inadequate" prenatal care	18.70%	19.60%	27.20%

- Non-participants may have a) chosen not to participate or b) not been effectively reached by the program
- What could explain less prenatal care for non-participants?

Other causal inference strategies

- What if we couldn't randomize?
- We will cover several *other* econometric strategies you can use
- Suppose we knew that there was a lot of variation in how much OB-GYNs serving Medicaid patients referred to NFP
  - Instrumental variables approach
- Suppose there was policy that allowed families who had a *prior* birth with an infant weighing <1500 grams to receive home based services during their current pregnancy?
  - Regression discontinuity approach
- Suppose we had a state where nurse home visiting was rolled out in some districts but not in others
  - Difference in differences approach

Approach to Econometrics in PHS2000B

- Learning objectives
  - To understand the objectives of econometrics and how it differs from other causal inference approaches
  - To gain exposure to some of the econometric methods being used most often in public health research
    - Focus on intuition
    - Understand how to interpret results
    - Get some exposure to how to do estimation
- You will not learn 1/10 of what you would learn in a full econometrics course
- Understanding of econometric techniques constantly evolving
- Using these tools in your research requires a lot of learning by doing
  - How do you set up your data, which method is the best fit, etc.

Learning to use a method

- Assumptions: It is important to know and state the assumptions of the method that are required to establish internal validity of estimates
  - Testable assumptions should be tested, and the method only used if these tests are passed
  - Untestable assumptions should be stated and the plausibility of these discussed
- Additional assumptions are required for valid confidence intervals, and these should also be stated and tested
- Analysis should reflect recent developments – awareness of advances in tests of assumption validity, estimation techniques, options for estimation, post estimation presentation of results

Debates between different disciplines

- Acceptability of untestable assumptions is to some extent a social norm
- Changes over time, and history matters – impact of key papers in field
- Real differences in assumptions between econometrics and epidemiology, plus differences in naming ideas
- May be difficult to export a method to a different discipline if the assumptions are novel in that field: but some overlap and convergence occurring

Break for Questions

Some social norms in Econometrics compared with Biostatistics and Epidemiology

- Use of linear probability model rather than logistic regression for binary outcomes
- Pervasive use of fixed effects but not random effects
- Less emphasis on data quality and reporting of construction of analysis samples
- Estimation of many models with different covariates or functional form as robustness checks
- Continuous explanatory variables rather than discrete groups (e.g., age)
- Estimation of marginal effects not relative risks or odds ratios

Break for Questions

Continuous development of methods

- The literature on each method advances rapidly over time
- New sources of bias always being discovered!
- Valid confidence intervals
  - Constructing valid confidence intervals requires additional assumption to those needed for having valid (consistent) estimates
  - Robust standard errors, clustering, small sample corrections (i.e. bootstrapping) relax additional assumptions needed for valid standard errors
- Different estimation techniques
  - Minimize errors, maximum likelihood, method of moments (see extra slides for more)
  - Can have slightly different assumptions
  - Monte Carlo studies of method performance
- Better algorithms for estimating coefficients
  - Shortcuts via transformation, convergence of iterative procedures
  - Faster computers allow brute force approach without assumptions to speed computation

Break for Questions

Additional slides

Estimator properties expressed mathematically

Define θ as the population parameter, $\hat{\theta}$ to be its estimate, and n to be the sample size

Unbiasedness: $E[\hat{\theta}] = \theta.$

Consistency: $P(|\hat{\theta}_n - \theta| > \delta) \to 0$ as $n \to 0$ for every $\delta > 0.$
Alternatively,
$\lim prob_{n\to\infty}[|\hat{\theta}_n - \theta| < \delta] = 1.$

Efficiency: $E\left[(\hat{\theta} - \theta)^2\right] = \min_{\hat{\lambda}} E\left[(\hat{\lambda} - \theta)^2\right].$

These definitions were presented in PHS 2000A, Lecture 3 as well.

I. Epi-Econ Dictionary

Developed by Emma Clarke, Doctoral Candidate, Global Health and Population Department, Harvard T. H. Chan School of Public Health

Omitted Variable Bias and Randomized Trials

Assume the **true *population* model** is given by:

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

Where we have demeaned the variables to remove the constant and we assume the error is i.i.d. random mean zero. However we estimate

$$Y = X_1\beta_1 + u$$

Which has omitted variable bias. We have

$$u = X_2\beta_2 + \varepsilon$$

Omitted Variable Bias and Randomized trials cont'd

The expected value of our estimate is given by:

$$\widehat{\beta_1} = (X_1{}'X_1)^{-1}X_1Y$$

$$\widehat{\beta_1} = (X_1{}'X_1)^{-1}X_1{}'(X_1\beta_1+X_2\beta_2 + \varepsilon)$$

and

$$E[\widehat{\beta_1}] = \beta_1 + \beta_2 E[(X_1{}'X_1)^{-1}(X_1{}'X_2)] + E[(X_1{}'X_1)^{-1}(X_1{}'\varepsilon)]$$

Since ε is random mean zero we have

$$E[\widehat{\beta_1}] = \beta_1 + \beta_2 E[(X_1{}'X_1)^{-1}(X_1{}'X_2)]$$

And there is omitted variable bias

However if X_1 is assigned randomly $E[(X_1{}'X_1)^{-1}(X_1{}'X_2)]$=0 and there is no bias

Note with random assignment the error term $u = X_2\beta_2 + \varepsilon$ from the "misspecifed equation" is uncorrelated with X_1

Term	Use in Economics	Use in Epidemiology
Selection bias	Economists think about selection bias as a problem arising from selection into the sample being studied.	Relationship between exposure and disease is different for those who participate in the study compared with all those who should have been theoretically eligible for the study. Epidemiologists think about selection bias as a problem arising from conditioning on common effects, or selection into the study.
Omitted variable bias	The bias that arises when a confounder is not included in a regression model.	Epidemiologists would typically use the term “unmeasured confounding” here instead.
Identified	A parameter is identified if different values give rise to different data distributions. It is not identified if two values generate the same distribution of the data	

Term	Use in Economics	Use in Epidemiology
Marginal effect	Marginal refers to a derivative (e.g., the marginal benefit is the additional amount of benefit that comes from consuming one additional unit of a good). Intuitively, you can think about "tweaking at the margins".	Marginal refers to an integral or a sum (e.g., the marginal treatment effect is the effect over the total population). Intuitively, you can think about the margins of a 2x2 table.
Endogenous	When a treatment variable is correlated with the error term in a linear regression creating bias. Usually used in the sense that (1) the dependent variable Y has a causal effect on X (simultaneity or reverse causation bias). Can also occur if there is (1) omitted variable bias; (2) measurement error.	The term is not typically used in Epidemiology.
Exogenous	Opposite of endogenous, uncorrelated with error term	This term is not typically used in Epidemiology

II. Comparing different estimators

Different estimators

- Minimize a function of unexplained residuals
  - Choose parameters to minimize residuals from a model
  - Requires an explicit model
  - Sum of squared residuals, sum of absolute residuals
- Maximum Likelihood
  - Choose parameters to maximize the likelihood of the observations given the parameter
  - Requires explicit model of how data are generated from parameters
- Method of Moments
  - Choose parameters to equate observed, sample moments with expected moments given parameters
  - Many possible moments to match

Minimizing Residuals

- Fit a model

$$y_i = \beta x_i + \varepsilon_i$$

- OLS Choose parameters to minimize

$$\sum_i \varepsilon_i^2 = \sum_i (y_i - \beta x_i)^2$$

Maximum Likelihood Principle

- Find the distribution (generally characterized by a vector θ) that maximizes the likelihood L of getting the collected sample data y

$$L(y, x, \theta) = f(y|x, \theta) = \prod_i f(y_i|x_i, \theta)$$

$$\log\big(L(y, x, \theta)\big) = \sum_i \log\big(f(y_i|x_i, \theta)\big)$$

$$\hat{\theta} = \underset{\theta \in \Theta}{a\text{rgmax}}\, L(y, x, \theta) = \underset{\theta \in \Theta}{a\text{rgmax}} \log\big(L(y, x, \theta)\big)$$

- These ideas were presented in PHS 2000A, Lecture 5 as well

Generalized Method of Moments

- Take any moment that holds true in the population given the model and impose it on the sample
- Moment condition: theoretical moment = sample moment
- Moment condition mathematically: $E[f(Y,X)|\theta] = \overline{f(Y,X)}$
- Solve this equation for θ

Coin Toss – is it a fair coin?: Estimator Comparisons

- Coin flip n tosses
- Outcome $y_i = \begin{cases} 1 \; if \;\; head \\ 0 \; if \;\; tail \end{cases}$
- Observe k heads
- We want probability of a head
- OLS estimate $y_i = \alpha + \mu_i$

$$\hat{\alpha}_{OLS} = (1'1)^{-1}1'y = n^{-1}\sum_i y_i = k/n$$

Coin Toss: Maximum Likelihood

- Maximum likelihood estimate

$$\underset{\alpha}{Argmax} \log \prod_i p(y_i|\alpha)$$

$$= \underset{\alpha}{Arg\max} \sum_i \log p(y_i|\alpha)$$

$$= \underset{\alpha}{Argmax} \left(k \log \alpha + (n-k) \log(1-\alpha) \right)$$

first order condition for a max

$$\frac{d}{d\alpha} \left(k \log \alpha + (n-k) \log(1-\alpha) \right) = 0$$

$$\frac{k}{\alpha} - \frac{(n-k)}{1-\alpha} = 0$$

$$\hat{\alpha}_{ML} = k / n$$

Coin Toss: Method of Moments

- Method of moments equate model moment and observed moment
- Observed moment – proportion heads

$$\bar{y} = k / n$$

- Model moment – expected proportion heads

$$E(\bar{y}) = \frac{1}{n}\sum_i E(y_i) = \frac{1}{n} n\alpha = \alpha$$

- Equating moments

$$\hat{\alpha}_{MM} = k / n$$

OLS and Maximum Likelihood

$$Y = X'\beta + u$$

$$L(y, x, \beta) = f(y|\beta, X)$$

$$f(y|\beta, X) = f(u = y - x\beta)$$

Further, assume one explanatory variable no constant and errors are normally distributed

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-u^2/(2\sigma^2)}$$

$$f(y - \beta x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\beta x)^2/(2\sigma^2)}$$

OLS and Maximum Likelihood

$$L(y, x, \beta) = \sum_i \log f(y_i - \beta x_i) = -n \log(\sigma\sqrt{2\pi}) - \sum_i (y_i - \beta x_i)^2 / (2\sigma^2)$$

Maximizing with respect to β, we get:

$$\frac{d}{d\beta} L(y, x, \beta) = 0$$

$$\sum_i -2(y_i - \beta x_i) x_i / (2\sigma^2) = 0$$

$$\hat{\beta}_{ML} = \sum_i x_i y_i / x_i x_i = (X'X)^{-1} X'Y$$

OLS and Method of Moments

Take the sample moments one for each explanatory variable

$$Sample\ Moments: \bar{y} = \frac{\sum_i y_i}{n}, \quad \overline{x_j y} = \frac{\sum_i x_{ji} y_i}{n}$$

$$Sample\ moments\ observed \quad X'Y/n$$

$$Model\ moments\ E(X'Y)/n = X'X\beta/n = \beta X'X/n$$

$$Equate\ sample\ moments\ and\ model\ moments$$

$$X'Y = \beta X'X$$

$$\hat{\beta}_{mm} = (X'X)^{-1} X'Y$$

Some notes on different estimators

- In simple cases the three estimation approaches are equivalent. OLS is minimum sum or residuals, method of moments, and maximum likelihood estimator (for assumed normal errors)

- In more complex cases they may give different estimators
  - Taxi problem: suppose we know taxi's in a town are numbered 1 to N we observe one taxi numbered k. What is your estimate of N using Method of Moments and Maximum Likelihood? Actually used in WW2 where enemy tank with number k was captured.

- As well as 3 different types of estimator we also have different estimators depending on details – e.g. function of residuals to minimize, functional form of error term in max likelihood, which moments we equate in method of moments