Introduction to Econometrics

AECN 896-002

Outline

1. Ligistics
2. What is econometrics about?
3. Causality and Association
4. Endogeneity

Logistics

Instructors

- Instructor : Taro Mieno (Office: 209, E-mail: tmieno2@unl.edu)
- Teaching Assistant : Shunkei Kakimoto (E-mail: skakimoto3@huskers.unl.edu)

Goals of the course

- Learn modern introductory econometric theory
- Apply econometric theories to real economic problems
- Learn how to use statistical software (R) so you can conduct research independently (without technical help from your advisor)
  - manage data
  - visualize data
  - run regressions
  - interpret results

Text Books

Required:

Wooldridge, Jeffrey M. 2006. "Introductory Econometrics: A Modern Approach (5th edition)." Mason, OH: Thomson/South-Western.

Recommended

- Florian, Heiss. 2016 "Using R for Introductory Econometrics." CreateSpace Independent Publishing Platform. (free version available online here)

Course Schedule

- Lectures (MW): 3:00-4:30pm
- Lab sessions (F): 1:00-2:30pm

Course Website

Course Website

- Lecture Slides
- Assignments
- Final paper

Grading

- Problem sets (4 assignments): 50%
- Paper: 50%
  - Proposal: 5%
  - Final paper: 45%

Assignments

Problem sets

- Most questions are from the required text book
- Some questions come from what we cover in lab sessions

Rmarkdown to do and submit your problem sets

- You are required to present your R codes
- You learn how to compile your assignment with your R code written in a document using Rmarkdown , which will be covered in the second lab session

Assignments

Caution

- 2nd year students have answers to all the questions I will assign (I will use exactly the same problems because they are really good to learn econometrics)
- You are free to copy and paste (or rephrase) the answers for your assignment. I won't bother to try to tell if you have copied and pasted answers.
- However, you are simply doing dis-service to yourself by depriving yourself of learning opportunities
- Moreover, your lack of understanding of the material will be clearly manifested on your final paper (I am not at all shy of giving bad grades on the final paper)

Paper

In this assignment,

- you write
  - a paper proposal with in-class presentation (5 points)
  - a paper with a particular emphasis on econometric analysis using a real world data set (45 points)
- you are encouraged to use the data set you are using for your masters thesis (talk with your advisor)
- you need to ensure that you use a panel dataset
- No presentation of your final paper

Paper

Here is the time line of the paper assignment:

- March, 23 : identify a research topic and the data set you will be using, and get an approval from the instructor
- April, 1 : paper proposal
- May, 11 : final paper

Paper Proposal

Introduction

- clear identification of what you are trying to find out (research question)
- why the research question is worthwhile answering

Simple Model

- dependent variable (the variable to be explained)
- explanatory variable (variables to be explain)

Data Source

- where you get data

Final Paper

Introduction

- clear identification of what you are trying to find out (research question) [1 point]
- why the research question is worthwhile answering [1 point]

Data description

- the nature of the data with summary statistics table [1 point]
- visualize a few key variables in a meaningful way [3 points]

Final Paper

Econometric Methods:

the process of how you end up with the final econometric models and methods. [40 points (or more)]

- justification of your choice of independent variables
- potential endogeneity problems
- what did you do to address the endogeneity problems?
- justification of econometric model(s) and method(s)
- identify appropriate standard error estimation methods

Results, Discussions, and Conclusions:

- interpret and describe the results [2 points]
- implications of the results [1 point]
- conclusions [1 point]

What is econometrics about?

What econometrics is about

Econometrics:

Estimate quantitative relationships between variables

Examples:

- the impact of fertilizer on crop yield
- the impact of political campaign expenditure on voting outcomes
- the impact of education on wage

Steps in Econometric Analysis

- formulation of the question of interest (what are you trying to find out?)
- develop an economic model of the phenomenon you are interested in understanding (identify variables that matter)
- turn the economic model into an econometric model
- collect data
- estimate the model using econometrics
- test hypotheses

Step 2: Develop an economic model

Example: Job training and worker productivity

$$wage = f(educ, exper, training)$$

- $wage$: hourly wage
- $educ$: years of formal education
- $exper$: years of workforce experience
- $training$: weeks spent in job training

Note: Depending on questions you would like to answer, the economic model can (and should) be much more involved

Step 3: Develop an econometric model

$$wage = f(educ, exper, training)$$

The form of the function $f(\cdot)$ must be specified (almost always) before we can undertake an econometric analysis

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u$$

$\beta_0, \beta_1, \beta_2, \beta_3$

- are the parameters of the econometric model.
- describe the directions and strengths of the relationship between $wage$ and the factors used to determine $wage$ in the model

u

- is called error term
- includes ALL the other factors that can affect wage other than the included variables (like innate ability)

Step 4: Collect data

- survey
- websites
- experiment

Data types

Cross-sectional Data

- a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time
- the data on all units do not correspond to precisely the same time period
  - some families surveyed during different weeks within a year

Cross-sectional Data

```
##        wage educ exper female married
##   1:   3.10   11     2      1       0
##   2:   3.24   12    22      1       1
##   3:   3.00   11     2      0       0
##   4:   6.00    8    44      0       1
##   5:   5.30   12     7      0       1
##  ---
## 522:  15.00   16    14      1       1
## 523:   2.27   10     2      1       0
## 524:   4.67   15    13      0       1
## 525:  11.56   16     5      0       1
## 526:   3.50   14     5      1       0
```

Data types: Time-series Data

Time-series Data Observations on a variable or several variables over time

- corn price
- oil price

Note:

- The econometric frameworks necessary to analyze time series data are quite different from those for cross-sectional data
- We do NOT learn time-series econometric methods

Data types: Panel (Longitudinal) Data

Panel (Longitudinal) Data time series data for each cross-sectional member in the data set (same cross-sectional units are tracked over a given period of time)

Example

- wage data for individuals collected every five years over the past 30 years
- yearly GDP data for 60 countries over the past 10 years

Notes

- Panel data are much more common than they used to be
- Panel data econometric methods take advantage of the panel data structure

Data types: Panel (Longitudinal) Data

```
##       county year    crmrte   prbarr  prbpris
##    1:      1   81 0.0398849 0.289696 0.472222
##    2:      1   82 0.0383449 0.338111 0.506993
##    3:      1   83 0.0303048 0.330449 0.479705
##    4:      1   84 0.0347259 0.362525 0.520104
##    5:      1   85 0.0365730 0.325395 0.497059
##   ---
## 626:    197   83 0.0155747 0.226667 0.428571
## 627:    197   84 0.0136619 0.204188 0.372727
## 628:    197   85 0.0130857 0.180556 0.333333
## 629:    197   86 0.0128740 0.112676 0.244444
## 630:    197   87 0.0141928 0.207595 0.360825
```

Steps 5 and 6

This is what you learn for the next few months!!

- estimate the model using econometrics
- test hypothesis

Causality and Association

Causality and Association

Association

An association of two variables arise because either of or both variables affect the other variable

$$A \longleftrightarrow B$$

Association does not concern which affects which. This is what correlation coefficient measures.

Causality Causal effect is the impact of one variable on the other,

$$A \rightarrow B$$

Here, changes in A cause changes in B, not the other way around

Let's watch this interesting CM.

Claims made in the video

People who wear glasses are

- much smarter than those who don't
- more likely to pursue higher education
- 200% more likely to graduate college

For you to be convinced to buy glasses, these claims needs to be causal, not association:

- Does wearing glasses make you much smarter?
- Does wearing glasses make it more likely for you to pursue higher education?
- Does wearing glasses make it 200% more likely for you to graduate college?

However, this seems to be a more likely explanation of the association:

- One spends more time studying academic subjects
  - smarter (or knowledgeable) $\rightarrow$ pursue higher education and graduate college
  - worsened eyesight $\Rightarrow$ wear glasses

Important:

- We care about isolating causal effects, but not association
- Identifying association is super easy
- Identifying causal effects is extremely hard (this is what we tackle)

Endogeneity: Your Nemesis

Causality and Association

It is super easy to find an association of multiple variables, but it is incredibly hard to find a causal effect (at least in Economics)!!

Endogeneity

You are interested in the causal impact of fire fighters on the number of death tolls in fire events

fire event	death toll	# of firefighters deployed
1	10	20
2	0	3
3	5	10
4	3	5
5	50	50

Questions

- How are they associated?
- Can you say anything about the causal effect of fire fighters deployment on the number of death tolls?

What happened?

You ignored an important variable!!

fire event	death toll	# of firefighters deployed	scale of fire
1	10	20	20
2	0	3	5
3	5	10	20
4	3	5	10
5	50	50	100

Endogeneity Problem

Endogeneity (Definition):

Variables of interest are correlated with some unobservables (variables that cannot be observed or are missing) that have non-zero impacts on the variable that you want to explain

In the above example,

- variable of interest : the number of firefighters
- unobservables : the scale of fire events (and other factors)
- variable to explain : death toll

The model:

$$\text{death toll} = \alpha + \beta \, \# \text{ of fire fighters} + (\gamma \, \text{scale} + v)$$

Endogeneity Problem:

# of fire fighters is correlated with scale, which we ignored

Another example: education on wage

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u$$

What are unobservables in u that are likely to be correlated with $educ$?

An important unobservable

- innate ability $\rightarrow$ wage
- innate ability $\rightarrow$ education

Most of the time, you will be faced with endogeneity problems caused by at least one of the followings,

- omitted variables (the scale of fire events, innate ability)
- self-selection
- simultaneity
- measurement error

Central Question

How can we avoid or solve endogeneity problems?

How to deal with endogeneity?

- You have two opportunities to deal with endogeneity problems
  - at the design stage
  - at the regression stage (what you will learn in this course)
- Econometrics has evolved to address endogeneity problems at the regression stage because randomized experiments are infeasible most of the time
- How about econometrics and other fields of statistics: Statistics, Psychometrics, and Biometrics?

How to deal with endogeneity?

Field	Design	Estimation Method
Econometrics	not feasible (often)	intricate
Many other fields	feasible	relatively simple

Deal with endogneity at the design stage

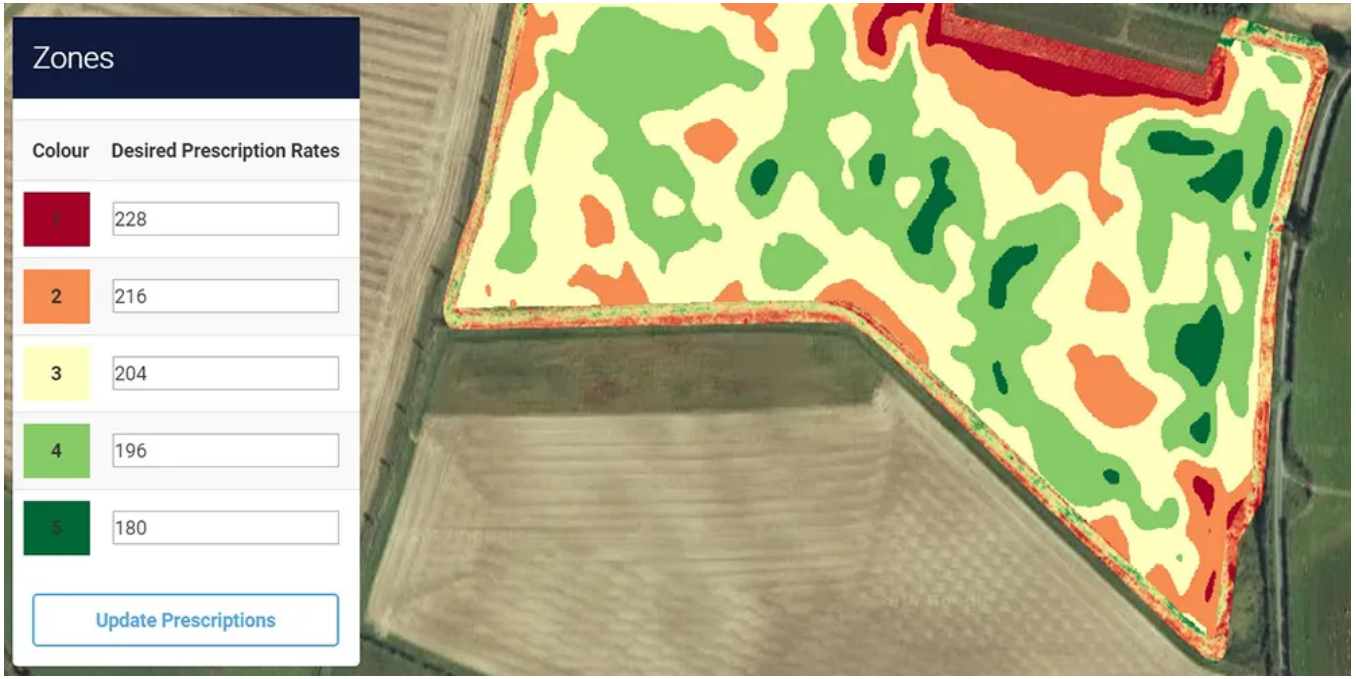
Randomized Experiments

- you have a liberty to determine the level of the variable of interest
- by randomizing the value of the variable of interest, you can effectively break the link (association) with whatever is included in the error term

The impact of fertilizer on corn yield (Non-Randomized)

Data:

Yield and nitrogen rate data obtained from a field that is managed by a farmer



The impact of fertilizer on corn yield (Non-Randomized)

Farmer

- decide nitrogen rate based on soil characteristics

Researcher

- soil characteristics is not observable, so it is in the error term

$$yield = \beta_0 + \beta_1 N + (\gamma SC + \mu)$$

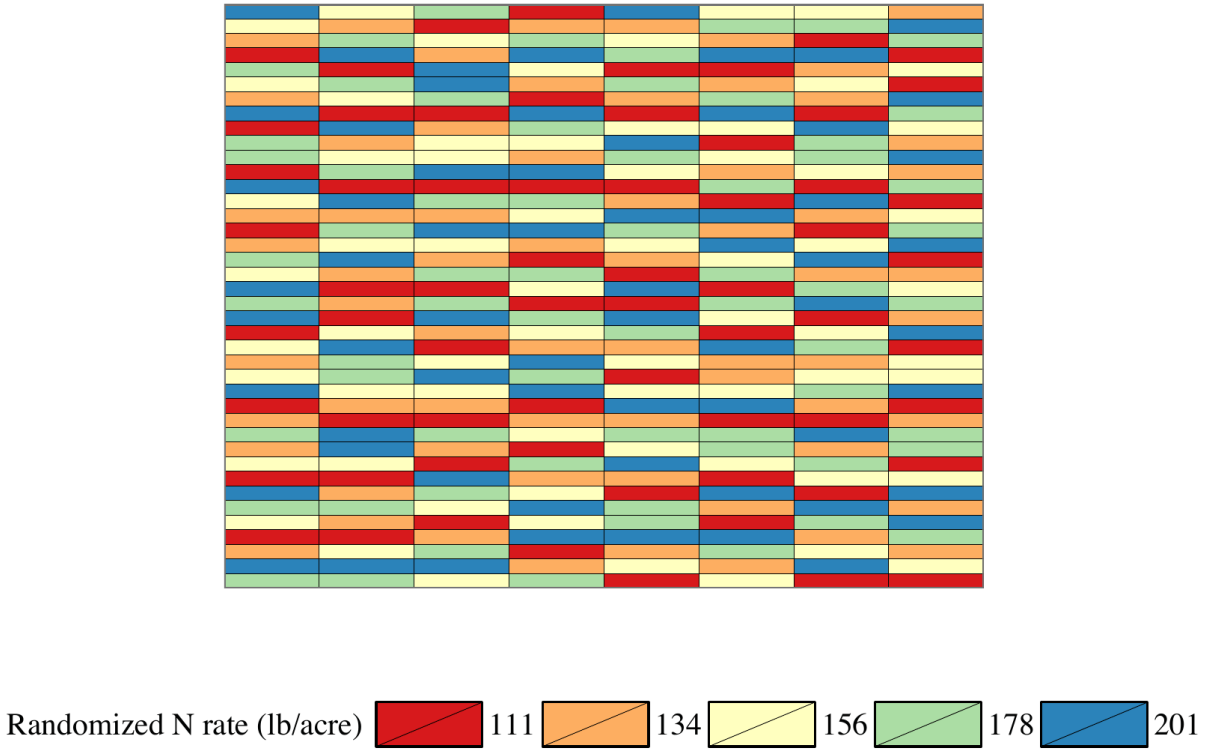
- N (nitrogen rate) and SC (soil characteristics) are correlated

The impact of fertilizer on corn yield (Non-Randomized)

Suppose the farmer applied more nitrogen to the area where its soil characteristics lead to higher corn yield

Question If the researcher estimate the model (which ignores soil characteristics), do you over- or under-estimate the impact of nitrogen rate on corn yield?

Randomized Experiments



Important Soil quality (in error term) is no longer correlated with N!!

Randomized Experiments on Education?

Randomized Experiment?:

Researchers determine randomly how much education subjects (people) can get?

Endogeneity Problem in Economics

- Economics is about understanding human behavior
- Almost always, you need to deal with endogeneity problem because people are `smart`: we make decisions based on available information (not just randomly) so that our decisions lead to good outcomes (whether our decisions turn out to be good or not is irrelevant)
  - how much education one get is determined based on their judgment of their own ability (not by rolling a dice)
  - how many fire fighters to be deployed was determined based on the scale of fire (not by rolling a dice)
  - how much nitrogen to apply based on soil characteristics (not by rolling a dice)
- If people are not smart and just roll a dice for their decision making, we would have much easier time identifying causal effects