Introduction to Econometrics

Andrea Velásquez

Department of Economics
University of Colorado Denver

January 17, 2018

Why Study Econometrics?

Econometrics applies statistical techniques and models to data in order to measure and interpret economic and social phenomena.

How is Econometrics different than Statistics?

- the difference is the “interpret” part of the previous bullet
- we seek to give meaning to the numbers generated by Statistics

For what purpose?

- To exploit information to reach goals like:
  - making better decisions (like how to invest your time and/or money)
  - improve the world (eradicate poverty, prevent disease, feed the hungry)
  - sell something, market something, get price right, predict future,
  - getting a job!! (don’t believe me?)

Why study Econometrics?

"*The ability to take data-to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it-that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complementary scarce factor is the ability to understand that data and extract value from it.*"

Why study Econometrics?

"*The ability to take data-to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it-that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complementary scarce factor is the ability to understand that data and extract value from it.*"

Hal Varian, Chief Economist, Google

What's the Catch?

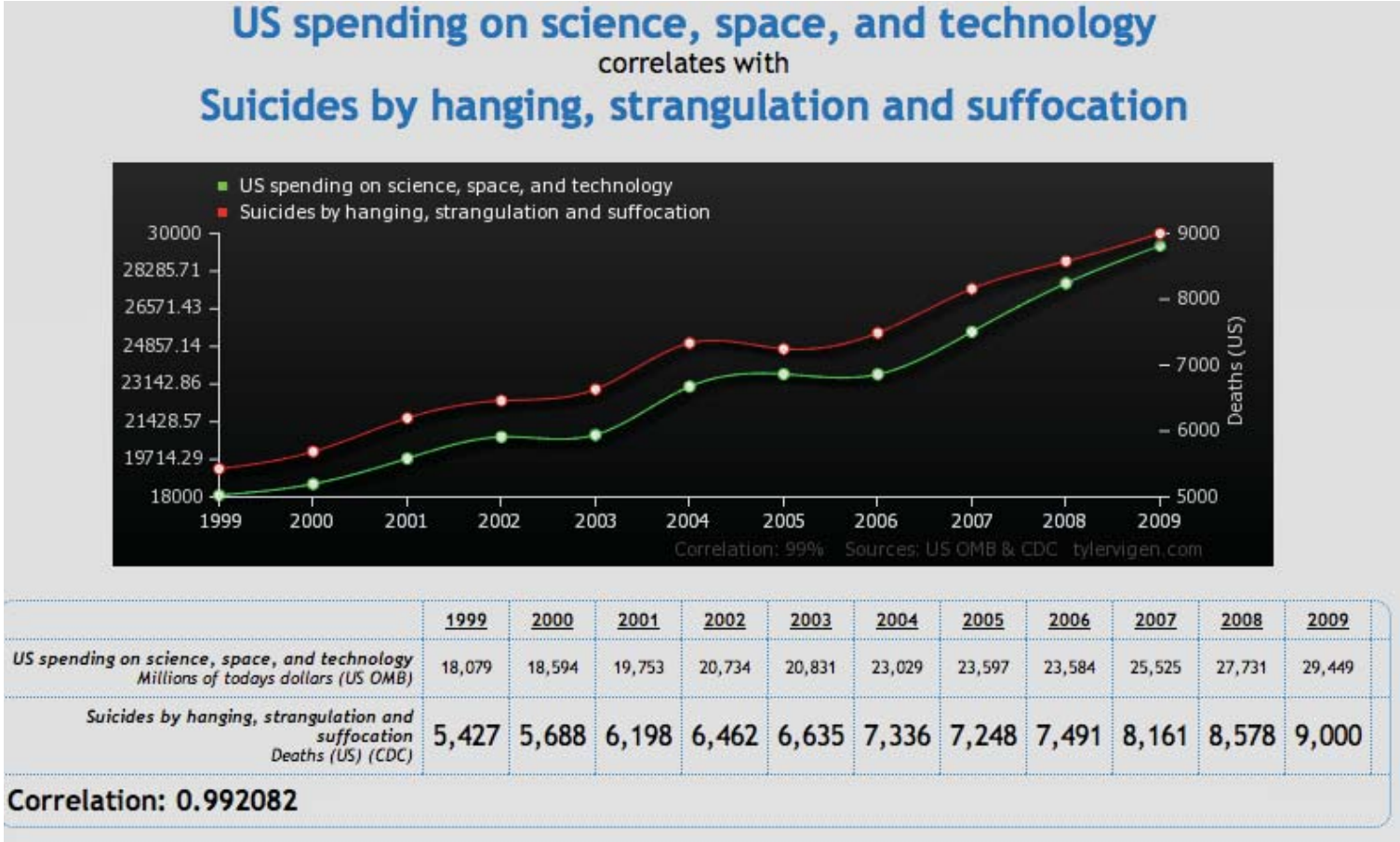
If econometrics is so great and valuable why doesn't everyone learn it?

Short answer.... because it is hard

Fundamental Problem 1 (Causation v Correlation):

- Does a change in X really *cause* a change in Y?
  Or do they just co-vary?
  - to evaluate policies and test theories we need to establish causation.
  - when the econometric methodology can uncover the true causal relationship we call the analysis **internally valid**
  - but in the real world, we observe relationships (correlations), while the causal mechanisms are often very difficult to distinguish.
  - why is correlation not enough?

Spurious Correlations



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Source: http://www.tylervigen.com

Causation vs. Correlation: Examples

We observe a *positive* relationship between crime and the number of police officers

- Is it because police officers create crime?
- Or is it because more police officers are assigned to more troublesome neighborhoods?

We observe a *negative* relationship between a newborn's birth weight and whether the mother drank while pregnant

- Does drinking while pregnant cause the *in utero* child to be born with lower birth weight?
- Or is it because the mothers who drink are inherently different and would have had lower birth weight anyway?

We observe that unemployed people who attend a job training program find a job faster

- Is it because the program helped them?
- Or is it because those who joined the program are the most motivated ones, and they would have found a job faster anyway?

So how do we establish causation?

Causation vs. Correlation

Ideally, we would like to observe the **same** person/country/neighborhood in **more than one** version of the world **simultaneously**, but we can't!

- e.g. same city with and without police, same mother drinking and without drinking while pregnant, same individual with and without job training...

What do people usually do as an alternative?
. . . compare "similar" people/countries/neighborhoods

- compare income at age 45 of person A who got job training to income at age 45 of person B who did not
- compare birth weight of children of a mother who drank while pregnant to birth weight of children of a mother who did not

Is this a good strategy?

Does this approach give us **internal validity**?
(i.e. can we make a causal statement using this approach?)

- almost surely NO. . . why not?

Causation vs. Correlation

What else can we do?

- **Experiments**:
  - e.g. unemployed are randomly assigned to attend a job training
- **The Good:**
  - random assignment means person A and person B are very very similar (if we gather enough people) . . .
    the only difference between them is the thing we are manipulating, called "the treatment" (e.g. job training)
  - done correctly, experiments provide **internal validity**. . . why?
  - because if the only difference between A and B is the treatment, then any difference in outcomes (income, health, eduction, etc) must have been **caused** by the treatment
- **The Bad:**
  - feasibility/ethics (e.g. can't force countries to have certain policies, can't randomly make pregnant women drink, etc.)
  - compliance (i.e. people will not always follow the rules)
  - so that's not the magic bullet
- **The Ugly:**
  - even if we can run an experiment...

What's the Catch?

- If econometrics is so great and valuable why doesn't everyone learn it?
- Short answer....because it is hard
- Fundamental Problem 2 (External Validity):
  - We want to make big, global statements!
    - but our data is on a subset of people, in one particular place, at one particular time (a sample)
    - so an ever-present concern is whether our conclusions can be extrapolated to other populations, in other regions, at other times
- Econometrics is the tool that will allow us to address these problems and answer questions that will improve the world we live in
  ...or at least give us a job!

Course Policies and Details: STATA

Statistical Analysis

- To combine the theory of econometrics with the practice of econometrics, you need to **apply** the theory
- To do that, you will need to use statistical software on a computer
- You may use whatever software you like ...but you should use STATA
- I will provide a STATA handout with helpful commands and STATA commands will be referenced in problem sets
- You can either purchase Stata ($45) by following the link in the syllabus
- Or use the computers in the Behavioral and Social Sciences Computer Lab in North Classroom 1009A

Evaluation

Quizzes: 20%

Evaluation: In-class quizzes (20%)

Will be given at random times either during the lecture or recitation to assess understanding of course concepts

Pop quizzes?! Why??

- This class is difficult and moves quickly
- To facilitate learning I expect interaction:
  you should ask me questions
  I will be asking all of you questions
- The quizzes help ensure everyone is engaged
- and encourages you to stay focused (which I will show you is quite important)

Your lowest quiz grade will be dropped. Should you miss a quiz, the missed quiz will count as a 0.

Evaluation

Quizzes: 20%

Problem Sets: 25%

Evaluation: Problem Sets (25%)

Will be assigned approximately bi-weekly, assigned two weeks before due date

Grading:

graded on a scale of 0-3:
0 (very poor grasp of the material)$\rightarrow$0%
1 (understanding needs improvement)$\rightarrow$60%
2 (very good understanding)$\rightarrow$85%
3 (excellent/perfect)$\rightarrow$100%

Guidance for problem sets:

Late homework will be given a grade of 0.
The worst grade on all your problem sets will be dropped.
They are due at the beginning of class on the assigned day.
Do not email your answers.
Working in groups is encouraged but **must hand in own PS** and note which students you worked with.
Answer each question clearly and concisely.
Clearly mark and interpret answers, do not just print STATA output (no need to provide STATA code).
TAs (and future employers) **require** typewritten answers.

Evaluation

Quizzes: 20%

Problem Sets: 25%

Midterm Exam: 20%

Final Exam: 35%

Evaluation: Exams

Midterm (20%):

in class on **Monday, March 5th**
there will be no make-up test
it is not possible to take the test early or late
if you have a valid excuse for missing the test, I will substitute your grade on the final for the midterm
you may not communicate with anyone inside or outside the classroom during the exam; this includes texting, emailing or any other form of electronic communication

Comprehensive Final(35%):

during finals week
same rules as midterm

You may bring one 8.5x11 page of formulae with you for reference during both the midterm and the final exam

Course Policies and Details: Points of Contact

This class is particularly difficult as it demands high level mathematical as well as analytical skills to succeed

My experience is that one of these two aspects of the course will take you out of your comfort zone

Always look very carefully at PS, quiz, and exam solutions

As mentioned before, one recipe for success to work together

There is no better colleague than your fellow student

Utilize each other's comparative advantage!

To reiterate: your classmates should be your first point of reference

Course Policies and Details: Points of Contact

For questions that can not be solved internally:

- you will have excellent support from your TAs:
  Larry Hamelin and Wayne Wohler
- they have office hours in the 1380 Lawrence Street Graduate Room LW-460,
  on **Wednesdays 10:00-12:00PM and Thursdays 10:00-12:00PM** or by appointment
- additionally I have office hours in LW-470Q on
  Mondays 10:00-12:00PM or by appointment
- if possible, e-mail your question before coming to office hours
- **Please** do not email me as a substitute for asking sybstantive questions. It is very hard to answer complex questions by email.

If there are any accommodations that you require, you must tell me with enough time to provide that service

If there is a predictable absence (religious, planned, etc.) you must tell me 2 weeks in advance

Course structure and logistics

Lectures (required)

Provide notes before class with key ideas and formulae
Final lecture notes will be available after the lecture

Why aren't lectures optional?

Reason 1

For each topic, my goal in lectures is to:

1. Build intuition
2. Formalize concepts
3. Apply concepts

Approach not used in any text book
Impossible to build econometric intuition through notes/books alone
Lecture notes act as required text... so study them outside class like you would any other textbook
Lectures notes and the recommended texts (found in the syllabus) are **complementary** to class not **substitutes**

Course structure and logistics

Why aren't lectures optional?

Reason 2

This class is essential to being an economist

It is the marker of what separates Econ majors from each other and other social science students in the marketplace

As such, is it not easy and grading is very strict.... and success in this class is highly correlated with attendance

Relationship Between Grades and Attendance

In 2012 we did not make attendance mandatory...

- What % of Duke students fail a class? $\rightarrow$ 0.5%
- % of Duke students failed this class in 2012 $\rightarrow$ 20%
  - conditional on failing this class in 2012:
    % who attended $\geq$**90**% of lectures $\rightarrow$ **0**%
    % who attended $\leq$**50**% of lectures $\rightarrow$ **80**%
- % of Duke students who got an A in this class in 2012 $\rightarrow$ 20%
  - conditional on getting an A in this class in 2012:
    % who attended $\geq$**90**% of lectures $\rightarrow$ **95**%
    % who attended $\leq$**50**% of lectures $\rightarrow$ **5**%

Relationship Between Grades and Attendance

What can we conclude about the relationship between grades and absence?

- Would you say they are strongly correlated? **Yes**
- In which direction? **Negatively**
- Can you say absences *caused* ECON 4811 grades at Duke in 2012? **No**
- Why not?
  - Omitted variables: additional factors correlated with both grades and attendance (organization, motivation, ambition, etc.)
  - Reverse causality: people doing badly stop coming to class

In 2013 we made attendance mandatory and ... failures went down (5%)

Relationship Between Grades and Attendance

How does this change on the course policies help with the problem of omitted variables (OV)?

- forced students with characteristics, which previously would have been correlated with absence and failure, to come to class
- thus we can compare students from 2012 that did not attend lecture to similar student in 2013 that did

How does our experiment help with the problem of reverse causality (RC)?

- now grades can not be causing attendance (since everyone has to come)
- thus potential reverse causality is no longer a problem

Can we talk now about a causal relationship between absences and grades? Is this experiment internally valid?

- non-trivial change in student composition
- my teaching improved or class got easier
- another outside factor or policy changed in 2013

Relationship Between Grades and Attendance

Let's assume our result is internally valid:
(Attendance caused ECON 4811 grades at Duke in 2012 and 2013)

Can we conclude that absences cause ECON 4811 grades (in classes I teach) in general? **Maybe**

In which cases would this **not** be true? (i.e. Not Externally Valid)

- Duke students are different than non-Duke students in a way that impacts the relationship between grades and attendance
- 2012 and 2013 are years in which some factor was uniquely different in a way that effects the relationship between grades and attendance
- Duke students in 2012 and 2013 are different than other students in other years in a way that effects the relationship between grades and attendance

Course Policies and Details: Logistics

- Lecture is required
- Why is lecture required? (with that analysis fresh in your brains)
  - Internal Validity:
    student composition didn't change btw 2012 and 2013
    sadly for me, my teaching didn't get that much better
    sadly for you, the class didn't get easier
  - External Validity:
    Duke ECON students aren't special
    2012 and 2013 were not special
    Duke ECON students in 2012 and 2013 were not special
- Which is a long way to say...

come to class... and be engaged!

Material to be covered

1. Data Analysis
   - Correlation
   - Linear regression model
2. Statistical theory
   - Distribution theory
   - Estimation and inference
3. Multivariate regression model
   - Estimation and inference
   - Interpretation
   - Models with discrete independent variables
   - Non-spherical errors
   - Models with discrete dependent variables
   - Omitted variables and sample selectivity
   - Instrumental variable estimation
   - Panel data methods

Data analysis: Quick review of some ideas

Statistics:
Seek to *summarize essential information* from a set of data

Example: We want to know whether this class is worth taking
(i.e. how much money does it add to future salary?)

Too expensive and tedious to collect data on all students (population)

We will seek to draw inferences about a population from a sample

Let's say we have two samples:
1) 100 CU Denver grads that took ECON 4811
2) 100 CU Denver grads that didn't take ECON 4811

And let's say we have the age 35 income for each of them (X_i)

Statistics: Review of Basic Concepts

A good place to start are **measures of central tendency**

Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}[X_1 + X_2 + ... + X_n]$$

- where n is the number of observations in the sample
- $\bar{X}$ is the mean of the income of people in that sample
- we can then compare the means from the two samples to get one statistic that gives us information about the value of this class
- what is problematic about means? Easily effected by outliers

Sample Median

- the middle most value in the sample (50% of the values are greater and 50% of the values are less)

Sample Mode

- the most frequent value in the sample

Statistics: Review of Basic Concepts

Also very useful is a **measure of dispersion**

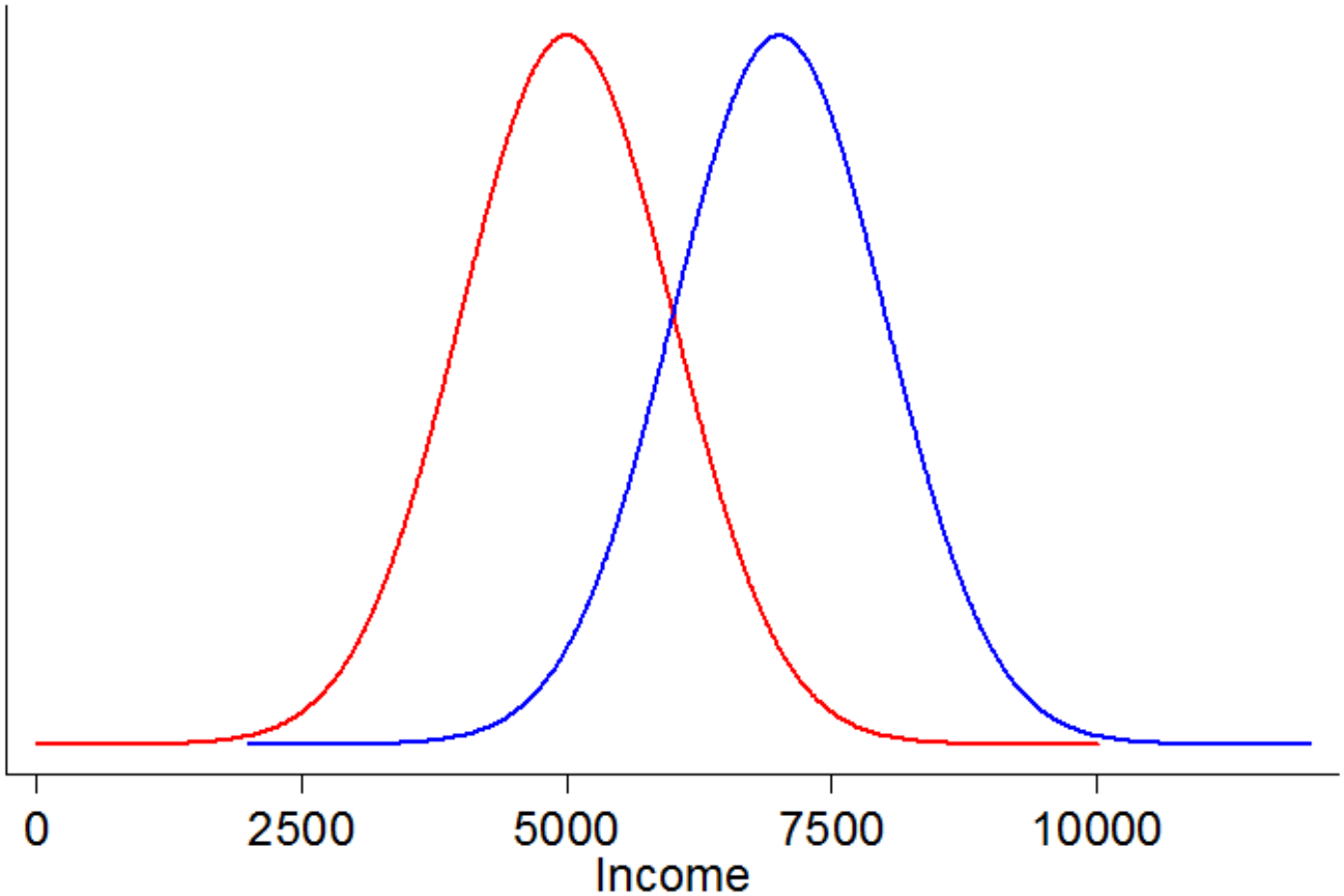
Sample Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- s^2 is an indicator of the dispersion of income for people in the sample
- why divide by (n-1)?
  1 less independent observation
  once you have n-1 of the observations and $\bar{X}$, the nth observation is completely determined

Would you rather be in the group with the higher measure of central tendency for income?

Which do you prefer?



Statistics: Review of Basic Concepts

Also very useful is a **measure of dispersion**

- Sample Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  - s^2 is an indicator of the dispersion of income for people in the sample
  - why divide by (n-1)?
    1 less independent observation
    once you have n-1 of the observations and $\bar{X}$, the nth observation is completely determined

Would you rather be in the group with the higher measure of central tendency for income? ....**YES!**

Would you rather be in the group with the higher measure of dispersion for income? ....**not so clear**

Which do you prefer?