ECON B253: INTRODUCTION TO ECONOMETRICS

Sebastian Anti, Assistant Professor of Economics
Bryn Mawr College

Lecture 1: Introduction to Econometrics

Overview

- Course Basics
- Syllabus and logistics
- What is this class about?
- The structure of econometrics research

Course Basics: Your Instructor

- **Name: Sebastian Anti**
  - "Sebastian" is fine
  - Office Hours: Monday 9:00 AM to 10:30 AM, Friday 12:00 PM to 1:30 PM
  - Pronouns: He/Him/His
  - Email: santi@brynmawr.edu (Only respond to email 8:30-5:00 PM M-F)
  - Website: https://www.sebastiananti.com/

- **Education:**
  - PhD, Applied Economics, University of Minnesota
  - MALD, Development Economics, Tufts University
  - BA, Oberlin College

- **Professional Background:**
  - Development economist
  - World Bank consultant
  - Peace Corps Volunteer, Cameroon 2008-2010
  - One Acre Fund Program Analyst, Kenya 2012-2014

Course Objectives

- **By the end of this class, you will be able to:**
  - Critically analyze econometric research with the appropriate intuition and fluency with foundational principles
  - Carry out econometric research with statistical software (Stata) in a professional capacity, advanced undergraduate course, or in graduate school
  - Pursue more advanced courses in econometrics at the graduate or advanced undergraduate levels

Expectations

- **Pre-requisites: ECON 105 (or equivalent) and a 200-level economics class**
  - Statistics, multivariable calculus, and linear algebra are not necessary
    - I will very occasionally refer to ideas from calculus and linear algebra but only to make connections for students with this background. It will never be material you are held accountable for
- **Prepared to work through (sometimes heavy) mathematical notation and reasoning**
- **A thorough review of required readings**
- **No lateness on problem sets**
- **Come to office hours for questions**

Requirements

- **Three Exams (Total 90% of final grade)**
  - Two Midterms (each 20% of final grade)
  - Final Exam (50% of final grade)
- **Five Problem Sets (Total 10% of final grade)**
- **Optional 6th Problem Set**
  - Can bring up weak midterm scores (more later)
- **Class participation is <u>not</u> a component of your score**
  - Although active participation in class is very encouraged!

Assignment Rules

- **Each exam is a scheduled sit-down exam**
  - The midterms will take 80 minutes
  - The final will take 120 minutes
- **For each exam you are allowed to bring in one 4×6 inch index card.**
  - You can write whatever you want on this card (front and back)
- **Problem Sets are done individually**
  - You may work with anyone you wish, but write your names of your collaborators on your write up
  - You may consult any resource you like (Stata User's Manual, StackExchange, etc.)

Assignment Rules

- **Problem sets require you to submit your Stata code in a ".do" file**
  - A .do file is the file that contains code that Stata can read
  - You may **collaborate** with others on your code, but you may not copy and paste code from another student and submit it as your own
  - Submit this on Moodle

Grading on Exams

- I generally start with an amount of points per question and take off points according to a schedule provided in your syllabus
- Multiple choice and true/false questions will have points attributed to them beforehand
- Do not neglect to answer questions
- Do not provide "non-answer" answers

Grading on Five Problem Sets

- **Graded on a Zero, Check minus, check, check plus basis:**
  - **Zero:** failed to turn in, mostly incorrect or incomplete, little skill development shown. No credit.
  - **Check minus:** Semi-complete. Limited ability with concepts demonstrated. Limited skill building. 50% in the grade book.
  - **Check:** Complete set of answers. Mostly correct. Adequate ability demonstrated. 100% in the grade book.
  - **Check Plus:** Complete set of answers: Almost all correct. Exceptional work. 100% in the grade book.

Feedback

- **I will highlight the general location of mistakes in your work so you know what to focus on when you go to look at the answer key**
- **I will not provide much feedback beyond this**
- **It is <u>your responsibility</u> to study the suggested answers that I provide to the problem sets to correct your errors**
  - You have the following resources
    - Me in office hours
    - Review sessions before midterms
    - Your peers
    - Optional TA sessions (scheduling in progress)

A Note on the Problem Sets

- **These are long**
  - Do not let them wait until the last minute
- **I recommend developing semester-long collaborative relationships with your peers in class**
- **If you hope to build skills in Stata and statistical computing, this is the main training this class offers**
- **We'll have an orientation to Stata on Thursday**
  - We'll be in Dalton's Computer Lab (Dalton 20) for this class

Optional 6th Problem Set

- **For students that want to improve their midterm scores before the midterm**
- **Optional for everyone**
  - Even if you did well on the first two midterms, this is good extra practice
- **If completed, I will recalculate your midterm grade before the final with your first two midterms receiving a 70% weight, and your score on the 6th problem set with a 30% weight**

Organization of Material

- **Organized around 7 Sections:**
  1. Introduction to Econometrics and Statistical Foundations
  2. Bivariate Regression
  3. Multivariate Regression
  4. Statistical Inference (of OLS coefficients)
  5. Binary Dependent Variables Models
  6. Methods for Causal Inference
  7. Real World Challenges when "Doing" Econometrics (If time permits)

Organization of Material

- **Note Well:**
  - "Sessions" in the syllabus do not align with a class session. Pay attention to what I say in class about what readings are coming up
    - This is because certain topics require more time than others
  - This schedule will change. Please pay attention to any changes that occur through the semester

Questions?

- **Any questions on the class organization or what you are responsible for?**
- **Be sure to read the syllabus carefully**
  - It has many details about the class we won't cover now but you are responsible for knowing
- **The rest of the class today will be content**

What is Econometrics?

- **Broadly Speaking: Econometrics is the methodology in economics for finding support for economic theories about how the world works**
  - It is a "way of knowing"
  - Effort to measure causal effects similar to lab experiments in the natural sciences (but with social science data)
- **More specifically, econometrics is the use of data to:**
  - Quantitatively measure the size of economic relationships and economic phenomena
  - Test economic theories
  - Predict economic outcomes

Difference between Theory and Econometrics

- **Theory offers:**
  - A framework for analysis that leads to conclusions and hypotheses
  - Conclusions are qualitative in nature
    - Example: If price elasticity of supply is greater than price elasticity of demand then consumers will bear the larger burden of an excise tax
- **Econometrics offers:**
  - Quantitative measurements of this relationship

Difference between Statistics and Econometrics

- **Statistics**
  - Frameworks for using sample data to make statements about populations (What's the difference?)
  - Summarizes the data
  - Little emphasis on theory driving the results and what it implies
- **Econometrics**
  - Applies statistical techniques to measure and inform our theoretical conception of the dynamics driving economic systems
  - Often more comfortable cautiously describing results as "causal" estimates (more in a bit on this)

What is Econometrics Used For?

- **Estimation**
  - **Demand Example:** How does quantity demanded for cars change as price increases?
  - **Policy Analysis Example:** What is the effect of a job-training program on participants' incomes?
- **Hypothesis Testing**
  - Does building a road to a village increase the probability of formal employment in rural India?
- **Forecasting**
  - What is the likelihood an individual becomes a returning customer based on prior consumption behavior?
  - What is US GDP likely to be in 2021?

Reasons to Study Econometrics

- **Helps understand economic systems**
- **Results of econometrics research used heavily in policy making**
  - Econometrics is used to measure the effect (the relationship) of policy interventions and outcomes like income, health, etc.
    - Example: What is the effect of a job training program on the income of participants?
- **Avoid being misled by research findings**
  - Econometrics commonly misinterpreted, or analysists intentionally overstate their findings
  - Technical nature often makes results opaque

Generic Econometrics Research Question

Most econometrics questions have a simple form:

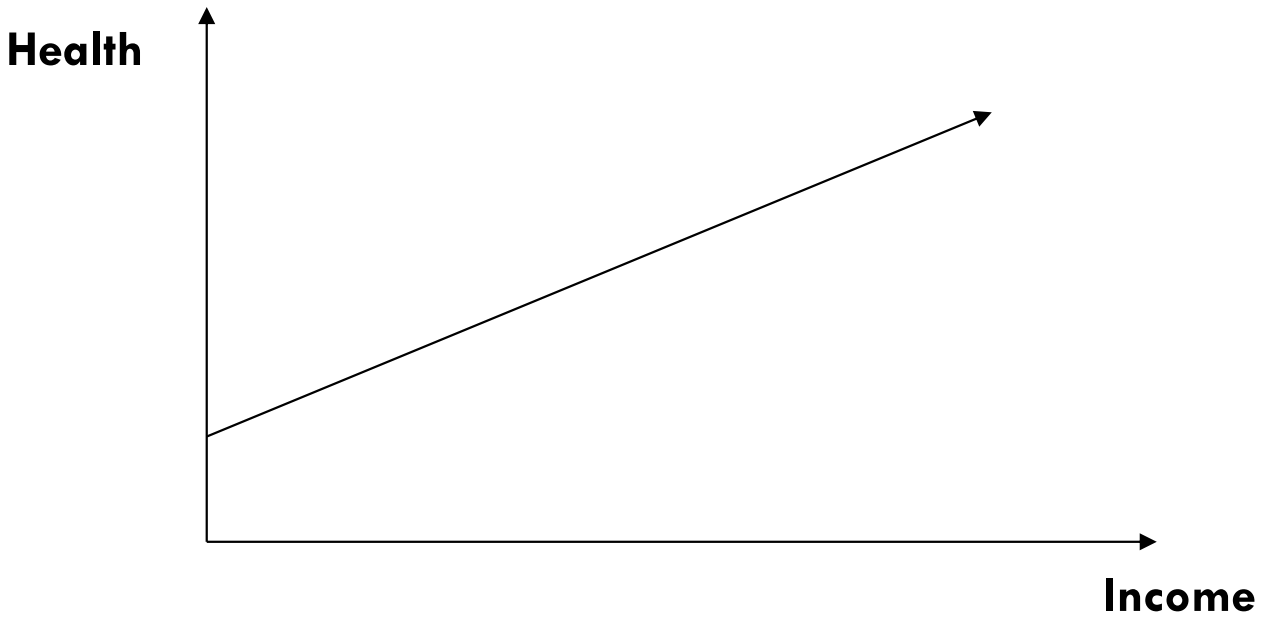
What (X) changes what (Y), and how much?

- **Examples:**
  - Income (X) and health (Y),
  - Price (X) and demand (Y),
  - Interest rate (X) and investment (Y).

Example: Income and Health

- **Theory, a mathematical model, suggests an exact relationship:**

$$Health = \beta_0 + \beta_1 Income \quad (1)$$



Slope of the Line Quantifies the Relationship

- **If we know the slope we've quantified the relationship!**
  - Slope describes the change in Y as X changes
- **But…**
  - You never actually see this line
  - Instead you have data from people on their health and income

Data Example

Tracking the individual respondent

Some measure of health status

A measure of income

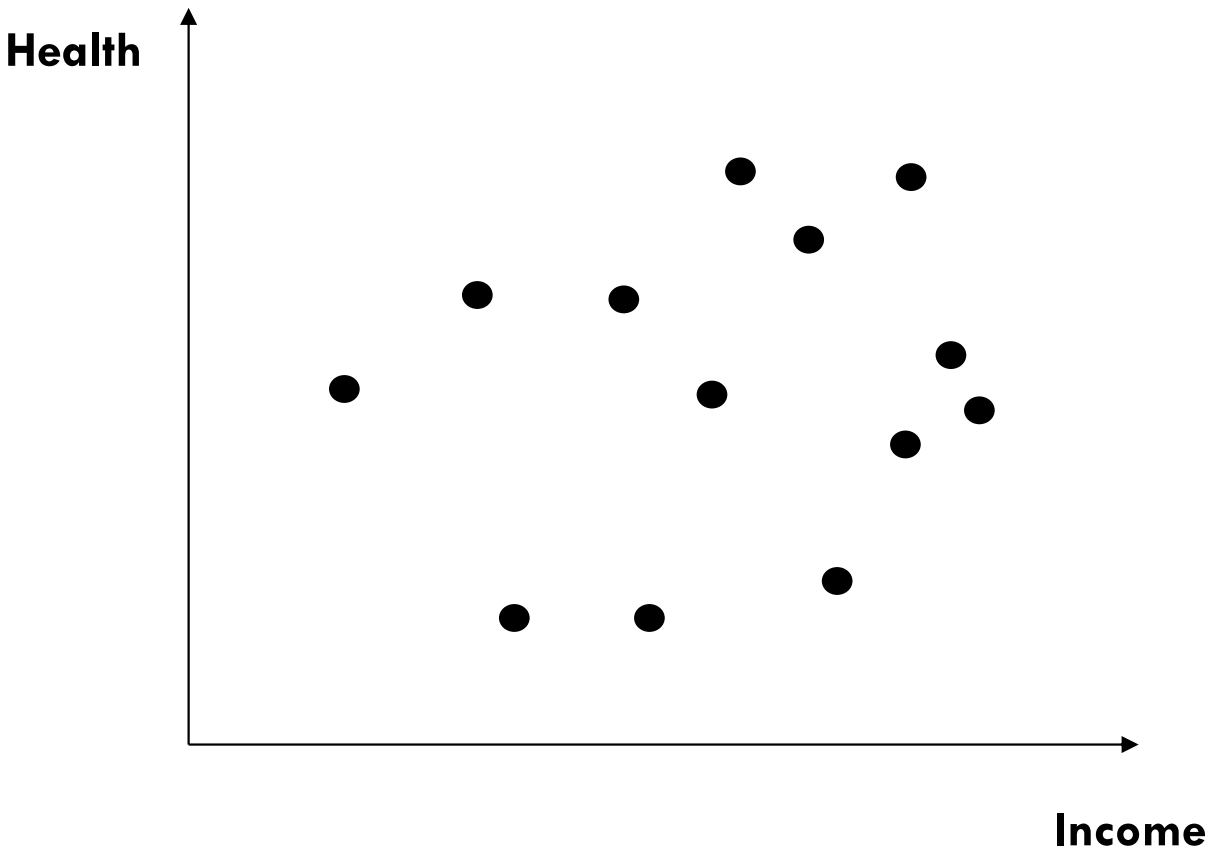
Individual ID	Self-Reported Health Status (1 = Bad, 10 = Perfect)	Annual Income
001	1	231,000 USD
002	3	192,000 USD
003	4	112,000 USD
004	1	56,000 USD
005	2	28, 000 USD
006	6	78,000 USD
007	10	20,000 USD

Data is stored in a spreadsheet (technically a matrix)

Each row (also called a "record") contains observations for one individual

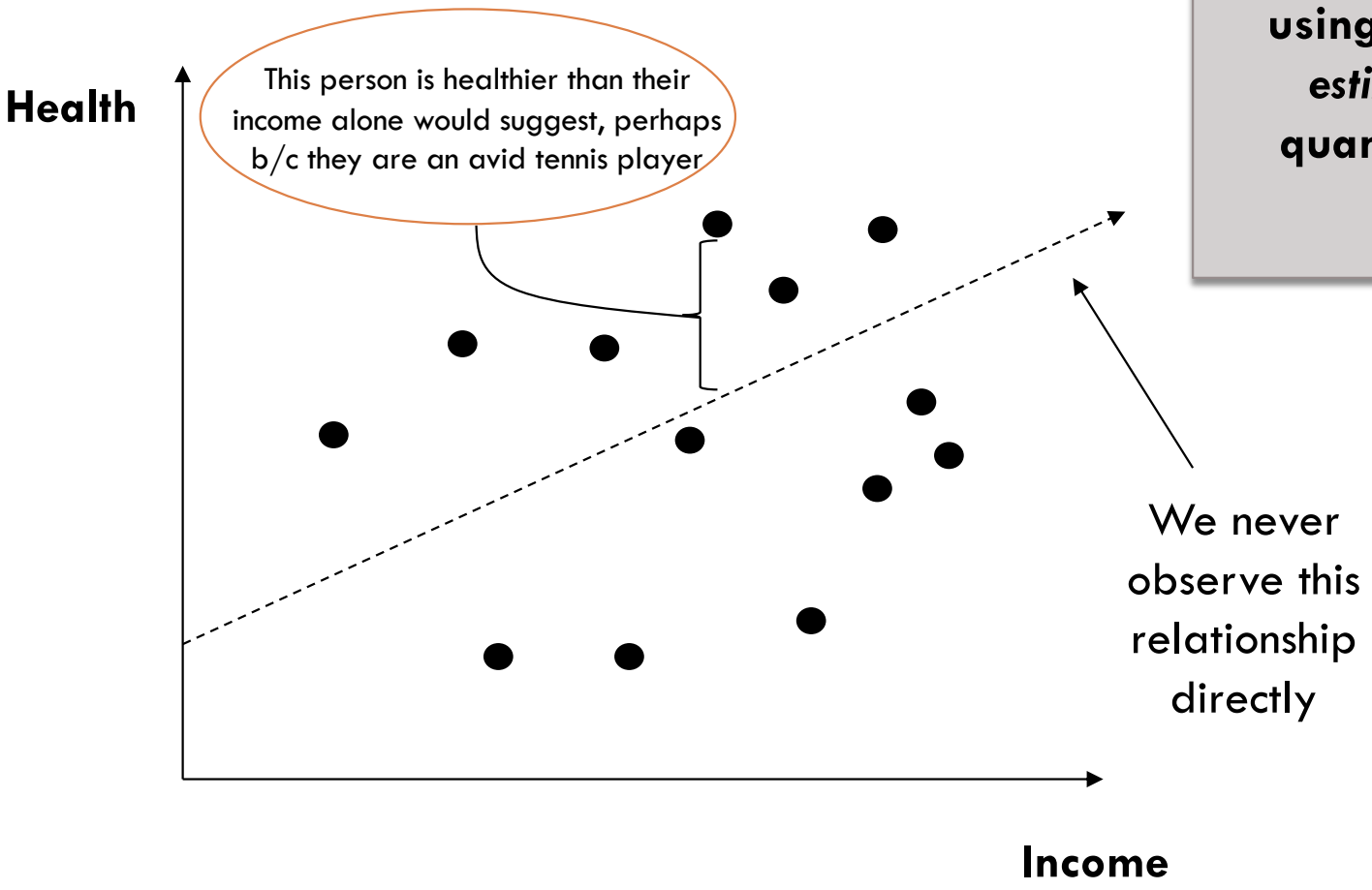
Each column (also called a "field") contains a particular variable

Data are Messy



Econometrics is the process of using that messy data to *estimate* this line (the quantified relationship)

Data are Messy



Econometrics is the process of using that messy data to *estimate* this line (the quantified relationship)

Steps in Conducting Econometric Research

1. Identify a **relationship or effect** you want to measure
2. Develop an **econometric model** of the relationship
3. Gather **data** measuring the variables in the model
4. **Estimate** the model
5. Perform **statistical analysis** of the estimates
6. Evaluate **potential for bias** in your estimates
7. **Interpret results,** taking into account statistical inference, potential bias, and economic significance of your results

Relationship or Effect

- Income's relationship to one's health?
- Education's relationship to one's wage?
- Participation in a job training program's relationship to whether one is employed?
- The relationship between rainfall and social conflict?

- What effects might we hypothesize for these?

Is it a Question of Correlation or Causality?

- **Correlations can be interesting, but scientific research is mostly interested in causal effects of X on Y**
- **Goal of econometric research is to produce the best possible estimate of how a change in X causes a change in Y**
- **Most often, a true causal estimate is elusive and requires many caveats**
  - This is true even with good data and a nicely scoped research question
    - Important to know what the analysis is really offering before citing it or claiming it as evidence for something

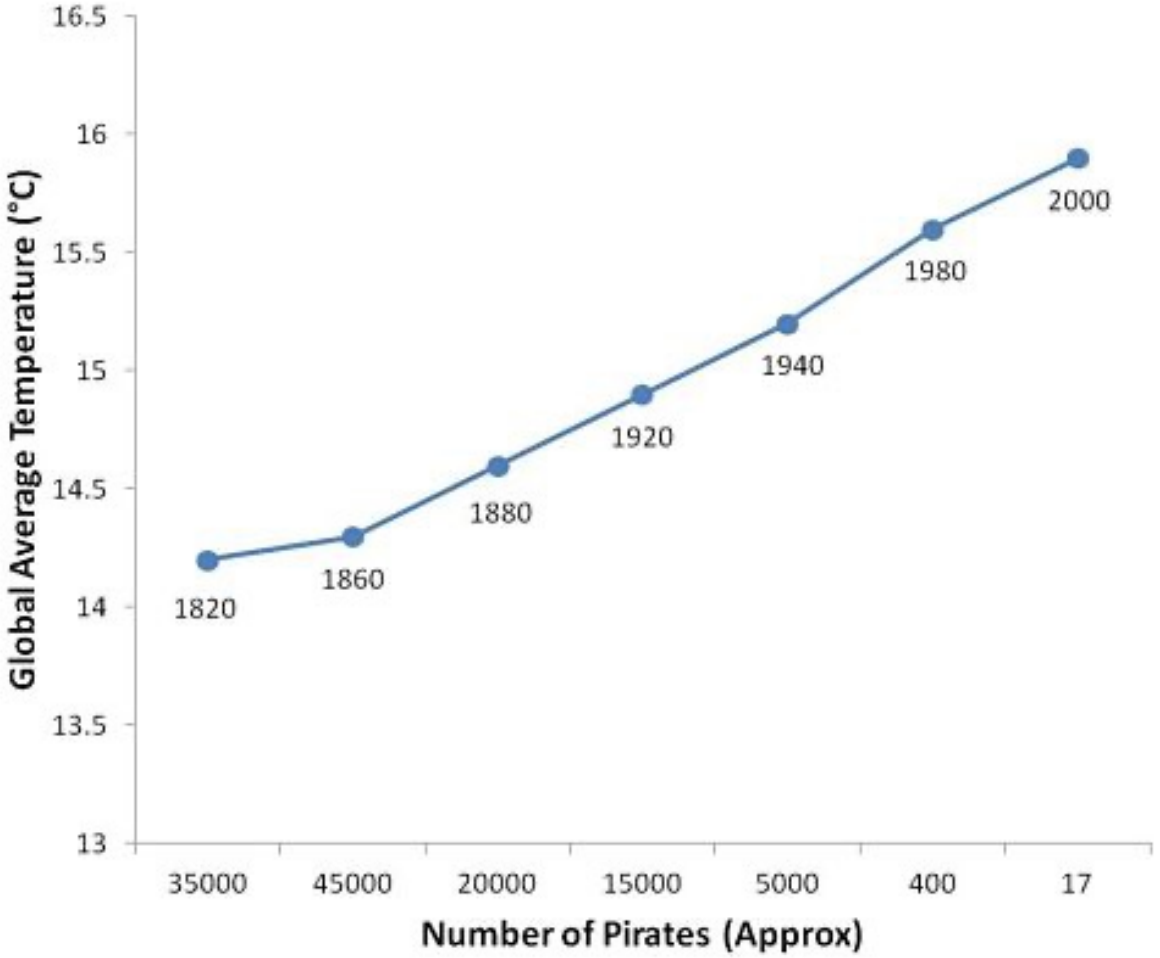
Causality versus Correlation in Detail

- **Correlation: Allows one to say "X and Y move together and are somehow related to each other"**
  - We'll learn about common measures of this
- **Causality ($X \longrightarrow Y$): Allows one to say "a change in X causes a change in Y"**
- **It's much much easier to show correlation than causality with data**
  - BUT causality is a lot more important when trying to understand how a complex system (like an economy) works

Not all Correlations are Causal

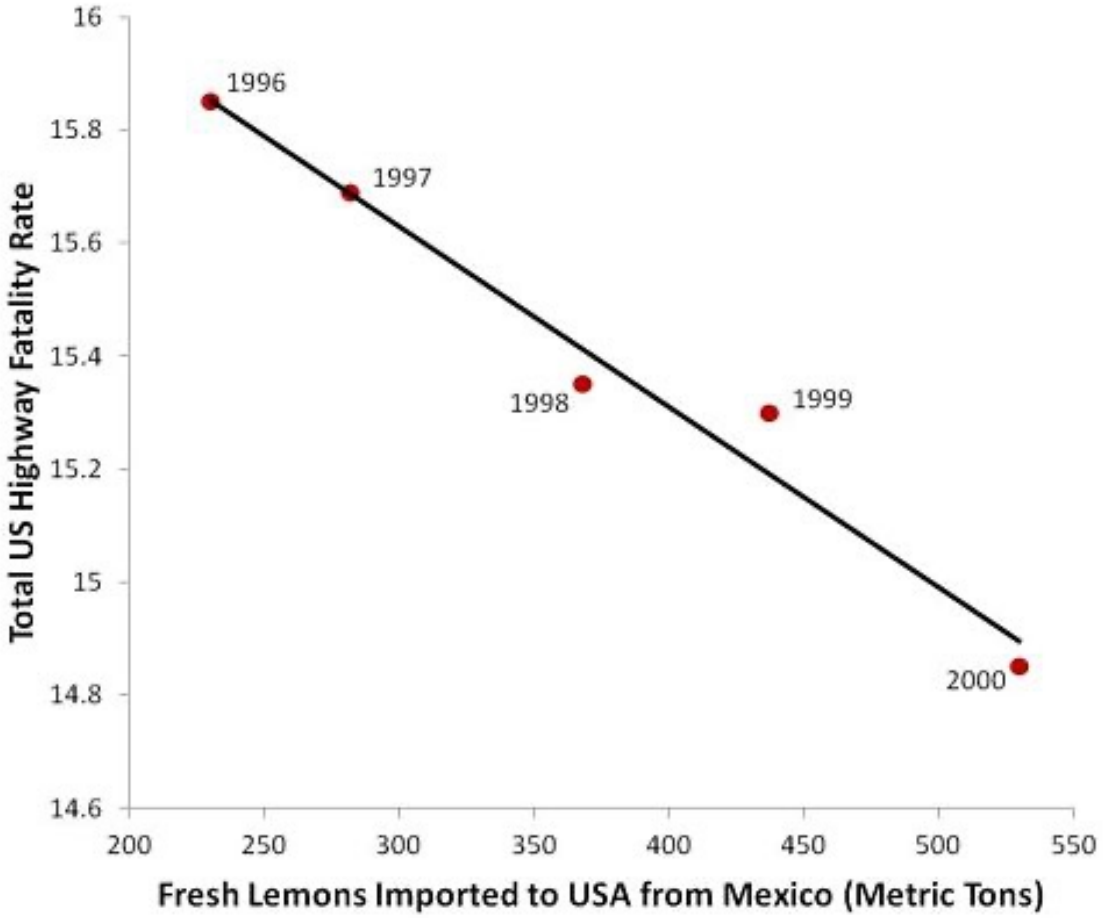
- **Why might X and Y be correlated?**
- **In others words, why might measurements of X and Y move together and have a relationship?**
  - **Many Reasons:**
    - X causes Y (causality)
    - Y causes X (reverse causality)
    - X and Y happen at the same time (simultaneity)
    - Z causes both X and Y (omitted variable)
    - Pure coincidence (spurious correlation)

Example: As the planet gets warmer, there are fewer pirates



- **Describe the chart. Does it mean that increasing global average temperature is causing fewer pirates?**
- **What do you think is driving this correlation?**

Example: As Mexican Lemon Imports Increase, Highway Fatalities in the US Decline



- **Pretty impressive correlation right?**
- **Anyone sold this is a causal relationship?**

Steps in Conducting Econometric Research

1. Identify a **relationship or effect** you want to measure
2. Develop an **econometric model** of the relationship
3. Gather **data** measuring the variables in the model
4. **Estimate** the model
5. Perform **statistical analysis** of the estimates
6. Evaluate **potential for bias** in your estimates
7. **Interpret results,** taking into account statistical inference, potential bias, and economic significance of your results

Develop an Econometric Model of the Relationship

- **An econometric model is an equation describing how X and all other factors determine Y:**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \quad \forall i \; where \qquad (2)$$

"for all"

Y = The variable being acted upon in the relationship of interest

X = The variable that may have a causal effect on Y that we can measure

Z = Another variable or group of variables that determine Y (there can by many. This equation just has one)

u = An "error" term

βs = "Parameters" that describe the relationship between X and Y and Z and Y

Develop an Econometric Model of the Relationship

Econometric model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \;\; \forall \, i \qquad (2)$$

- **The i subscript tells us this relationship is assumed true for individuals in the population from $i = 1$ to $i = N$**
  - In other words, it doesn't matter which individual we look at, this relationship should hold
- **Y for any individual is determined by their level of X and Z and u**
- **This relationship takes a specific functional form (linear in this case)**

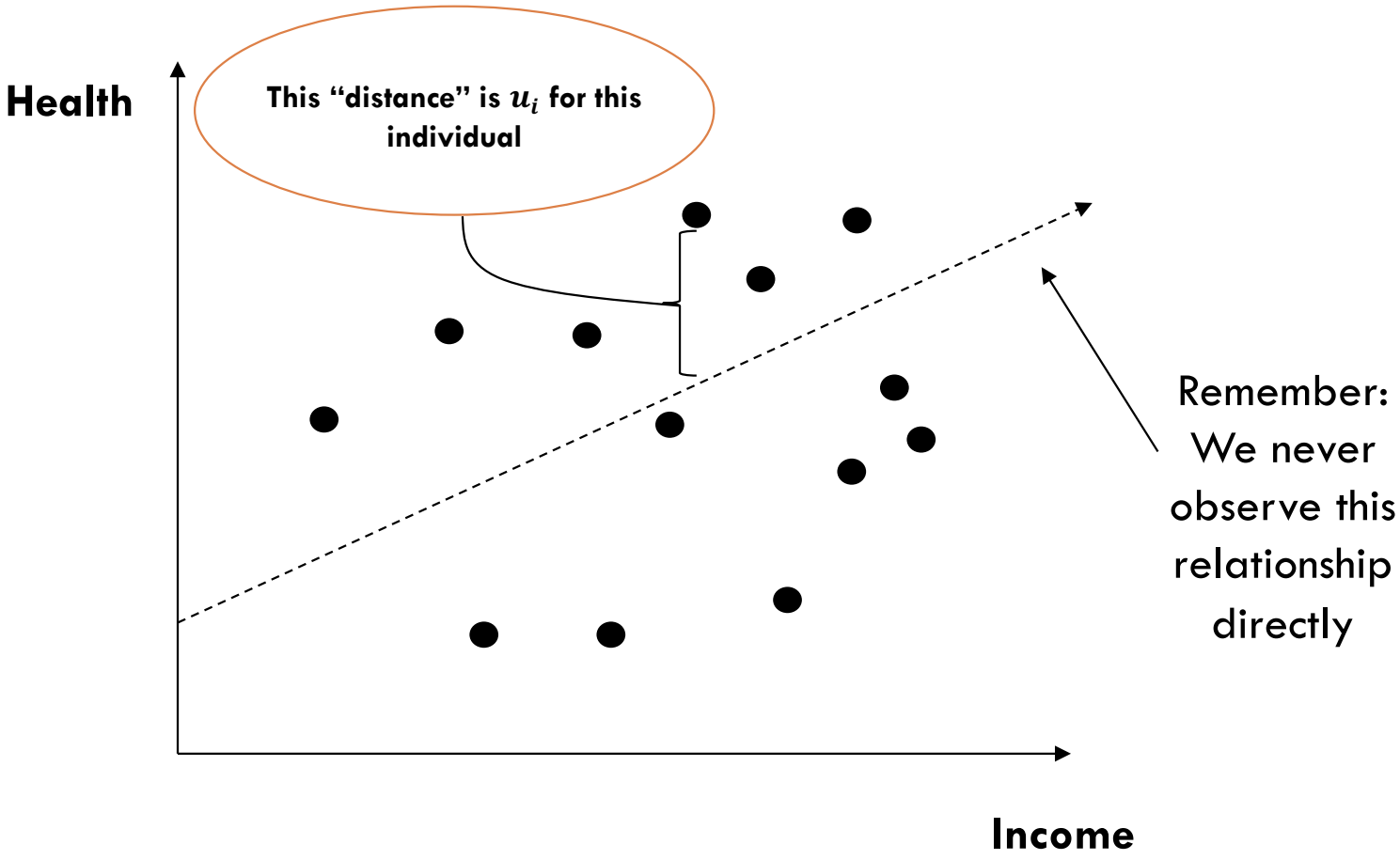
Develop an Econometric Model of the Relationship

Econometric model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \quad \forall\, i \qquad (2)$$

- **The error term (u) tells us that although we may know an individual's X and Z, we still won't be able to predict Y exactly**
  - We will assume that if we look at a bunch of individuals with specific values of X and Z, the average u among them would be zero
  - Common to think about u as "noise" that causes some people to be above the average and others to be below
- **We'll talk much more about u later**

Data are Messy



Develop an Econometric Model of the Relationship

Econometric model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \;\; \forall \, i \qquad (2)$$

- **If the model is correct, we can measure the βs, the "parameters"**
  - β_1 is our "parameter of interest." The quantification of the relationship between our variable of interest X and our outcome variable Y, holding Z constant
- **With the parameters we can answer:**
  - If we hold Z constant and increase X by a unit, how does Y change?
  - As X increases what happens to average Y?

How are we doing?

Important Concerns

- **To get an accurate estimate of the parameter of interest and other parameters, the model must be a good reflection of the *true model***
  - The true model is the the real world set of variables generating Y
  - True model is usually not known in social science
- **Social science reasoning must be applied to decide on the functional form and the set of variables to include in the econometric model**
- **We must think carefully about what is in u**
  - Is it just "noise" or is it systematically relates to the variables accounted for in the model?

Steps in Conducting Econometric Research

1. Identify a **relationship or effect** you want to measure
2. Develop an **econometric model** of the relationship
3. Gather **data** measuring the variables in the model
4. **Estimate** the model
5. Perform **statistical analysis** of the estimates
6. Evaluate **potential for bias** in your estimates
7. **Interpret results,** taking into account statistical inference, potential bias, and economic significance of your results

Gather Data

- **Many types of data gathered from all sorts of contexts**
  - Data availability is getting better by the minute
- **Data determines which econometric method(s) is/are available for the analyst to use**
- **Common types of data**
  - **Cross sectional:** Measures collected from a random sample of individuals, firms, households, cities, states, or countries at a given point in time
  - **Time series:** Measures collected over many time periods for a single entity like a country or a company's stock
  - **Pooled cross sections:** Two or more cross sections of information from different samples at different points in time
  - **Panel:** A time series (often very short) of measures on each cross-sectional member of a data set

Data Sources

- **Data can come from all types of sources**
  - Survey data
  - Administrative data
  - Remote sensing data (satellite imagery)
  - Price data
- **Data can be at different "levels"**
  - Individual
  - Firm
  - County
  - State
  - Country
  - Markets

Steps in Conducting Econometric Research

1. Identify a **relationship or effect** you want to measure
2. Develop an **econometric model** of the relationship
3. Gather **data** measuring the variables in the model
4. **Estimate** the model
5. Perform **statistical analysis** of the estimates
6. Evaluate **potential for bias** in your estimates
7. **Interpret results,** taking into account statistical inference, potential bias, and economic significance of your results

Estimate the Model

- **We need to use the data to derive an estimate of the parameter(s)**
  - In other words, to know the characteristics of that line we can never see
- **Need to take a systematic approach**
  - We want an algorithm to "fit" the model to the data that we can apply in every situation
- **Most econometrics uses the "Ordinary Least Squares" (OLS) algorithm (also called an "estimator") to take the data and derive estimates of parameters**
  - We'll also briefly spend time looking at "Maximum Likelihood" estimators

Estimate the Model

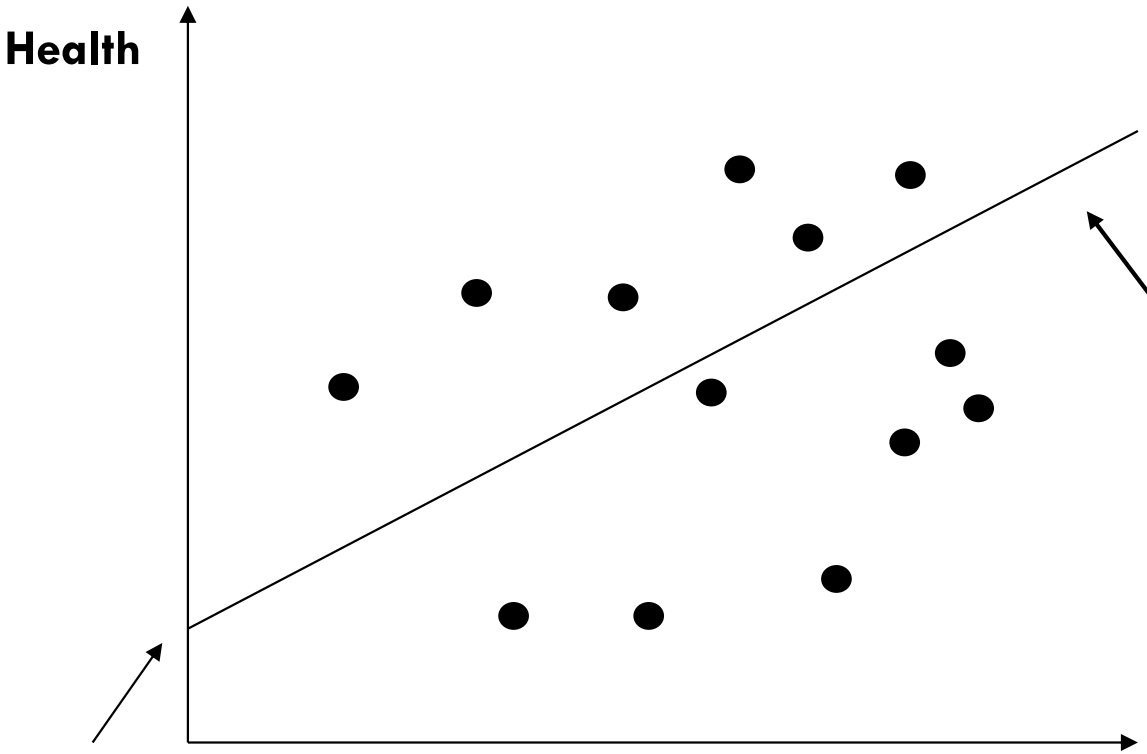
- **Let's consider the model,**

$$Health_i = \beta_0 + \beta_1 Income_i + u_i \qquad (3)$$

- **Assume that $Income$ is the only important determinant of $Health$, and u is not systematically related to income**
- **Also assume: for any level of $Income$, u is positive for some people and negative for others, but it averages out to 0 in the group**

Estimating the Model

$$Health_i = \beta_0 + \beta_1 Income_i + u_i$$



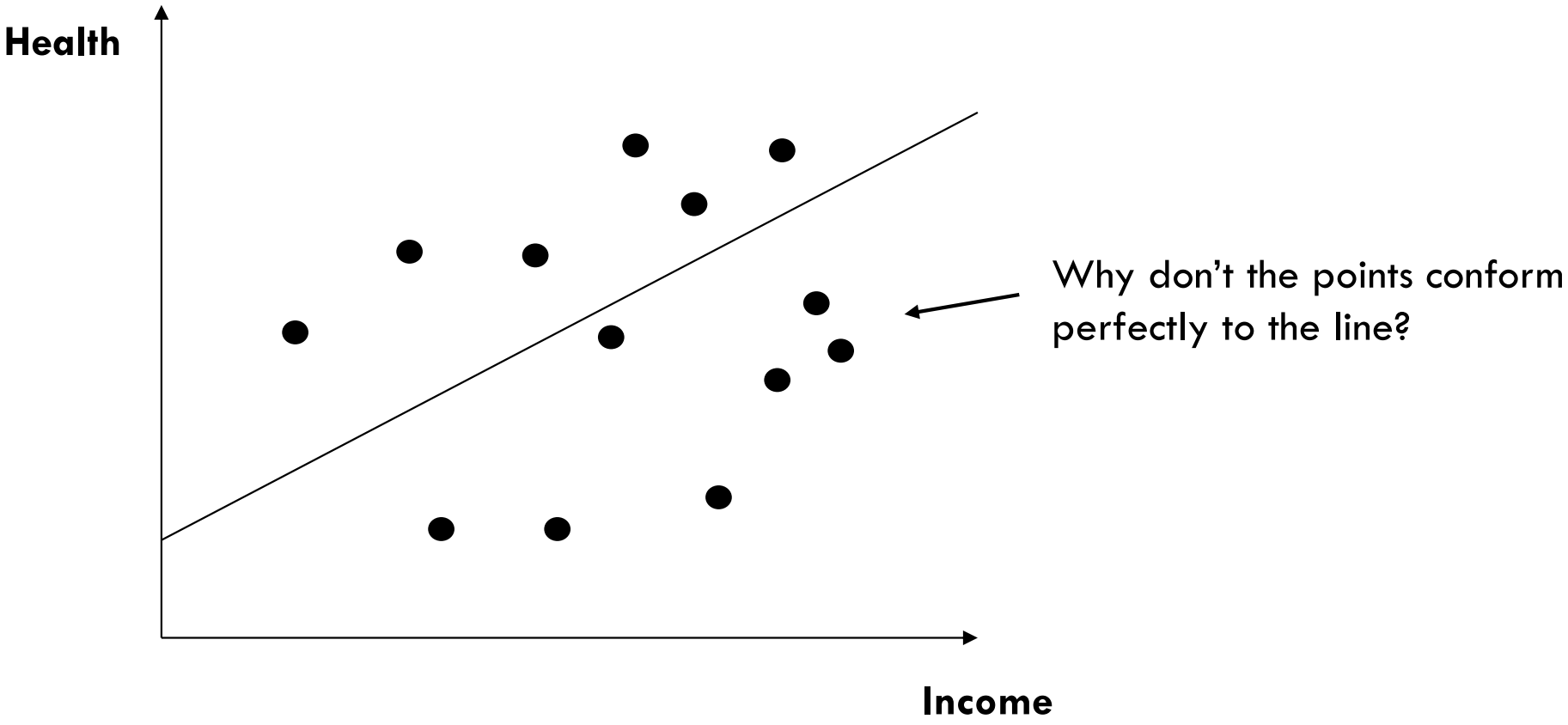
Relationship between health and income.

Estimated slope using OLS is $\hat{\beta}_1$. What does it tell us?

What does the intercept tell us?

Estimating the Model

$$Health_i = \beta_0 + \beta_1 Income_i + u_i$$



Estimating the Model

- **Understanding OLS in detail is critical because its properties influence the interpretation of the estimates**
  - It may feel like we are covering minutiae at points, but if you don't understand the behavior of OLS, you don't really understand econometrics
  - We'll develop notions of what it means to "best fit the data"
  - We'll also figure out the conditions under which the estimates we get will be "good estimates" of causal effects

Steps in Conducting Econometric Research

1. Identify a **relationship or effect** you want to measure
2. Develop an **econometric model** of the relationship
3. Gather **data** measuring the variables in the model
4. **Estimate** the model
5. Perform **statistical analysis** of the estimates
6. Evaluate **potential for bias** in your estimates
7. **Interpret results,** taking into account statistical inference, potential bias, and economic significance of your results

Statistical Analysis

- Data collection is expensive so we can never record information for the whole "population" of interest. We need to collect a random "sample"
- What if our sample is not representative of the population as a whole? We will never know whether it is or not.
- BUT statistics allows us to assess whether the data are sufficiently informative to give us confidence in our results
- When statistical analysis tells us that we can be reasonably confident that the estimates are not the product of just drawing a crazy non-representative sample, we say the estimates are "statistically significant"

Steps in Conducting Econometric Research

1. Identify a **relationship or effect** you want to measure
2. Develop an **econometric model** of the relationship
3. Gather **data** measuring the variables in the model
4. **Estimate** the model
5. Perform **statistical analysis** of the estimates
6. Evaluate **potential for bias** in your estimates
7. **Interpret results,** taking into account statistical inference, potential bias, and economic significance of your results

Evaluating "Bias"

- **Even if statistical analysis indicates we're measuring a real signal of a relationship (we have "statistical significance"), it doesn't mean our parameter estimates are correct measures of causal relationships**
  - Parameters could still just be indicative of reverse causality, or omitted variables, spurious correlation, or simultaneity
  - If we have reason to believe our measure is systematically incorrect, it is **"biased"**
    - **Note this is a technical term in econometrics and has a very particular meaning distinct from colloquial use**
- **We avoid these problems with experimental and quasi-experimental approaches (covered later in the semester)**

Steps in Conducting Econometric Research

1. Identify a **relationship or effect** you want to measure
2. Develop an **econometric model** of the relationship
3. Gather **data** measuring the variables in the model
4. **Estimate** the model
5. Perform **statistical analysis** of the estimates
6. Evaluate **potential for bias** in your estimates
7. **Interpret results,** taking into account statistical inference, potential bias, and economic significance of your results

Interpret Results