Lecture 1: Why Econometrics?

Pierre Biscaye

Fall 2022

Acknowledgement: Lecture material in this course relies heavily on materials produced by Professor Jeremy Magruder.

Econometrics applies statistical analysis to economic data

- We often want to examine relationships between two variables
  - Usually referred to as X and Y; for example, could look at relationships between:
  - Country: X=GDP/capita and Y=CO2 emissions
  - Individual: X=education years and Y=income
- In many cases want to develop *causal* interpretations
  - Does X=smoking *cause* an increased risk of Y=lung cancer?
  - Do X=school closures *cause* reductions in Y=women's employment?

Econometrics helps evaluate policy and test theory

- In some cases, want to evaluate the impact of policy on outcomes
  - Do mask mandates reduce the spread of COVID-19?
  - Does providing treated mosquito nets lead to higher human capital attainment?
- In others, may want to test a theory or hypothesis
  - Do higher minimum wages increase unemployment?
  - Does maternity leave increase women's labor force participation?
  - Does (lack of) access to credit prevent poor farmers from achieving high yields?

Why study Econometrics?

- Tools for understanding how to think critically about relationships between variables, particularly causal relationships
- Provides a broadly transferable lens to view the world
  - For example, will help you evaluate news reporting of scientific studies or government policies
- Economics wants to be a *science*: need the right tools to apply the scientific method

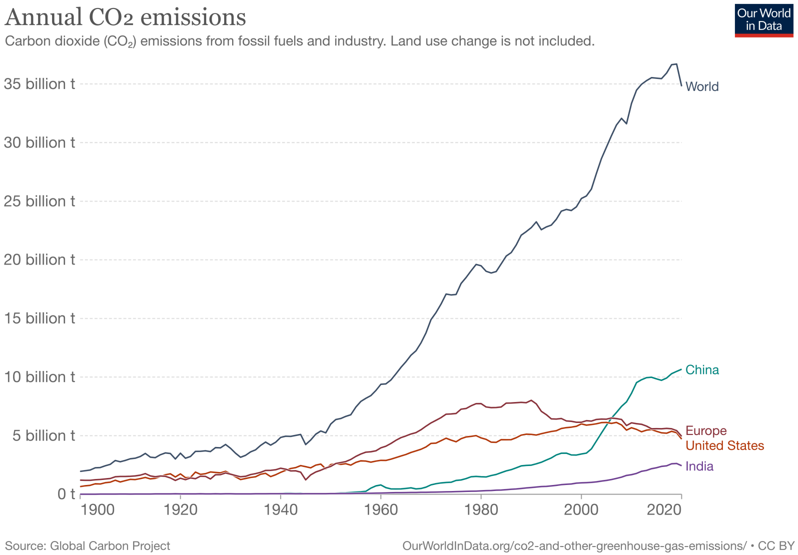
Why is Econometrics different from statistics and data science in other fields?

- Focus on causal inference rather than prediction
- Economics often wants *causal* answers
  - E.g., what is the effect of mandating masks on COVID-19 infections?
  - But, state governments did not adopt mask mandates at random $\Rightarrow$ how to identify the effect of mask mandates alone?
- Econometrics was built to ask: what can we say about the *causal* impact of policy when we only have *observational* data?
  - Occasionally we will have *experimental* data as in other sciences; causal inference is easy in this case
  - But a lot of the time, we will not. In this class, we'll cover some approaches to estimating causal effects outside of experimental settings.

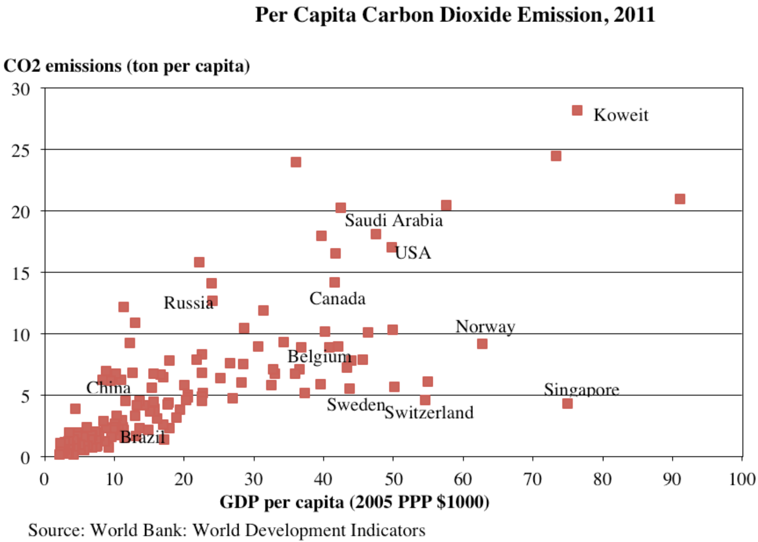
Example: GDP and CO_2 emissions

- Good News: most countries are becoming richer over time
- Bad News: what does this mean for CO_2 emissions and climate change?
- Policy need: how should carbon policy account for global economic growth?
- What do the data say?

CO2 emissions over time



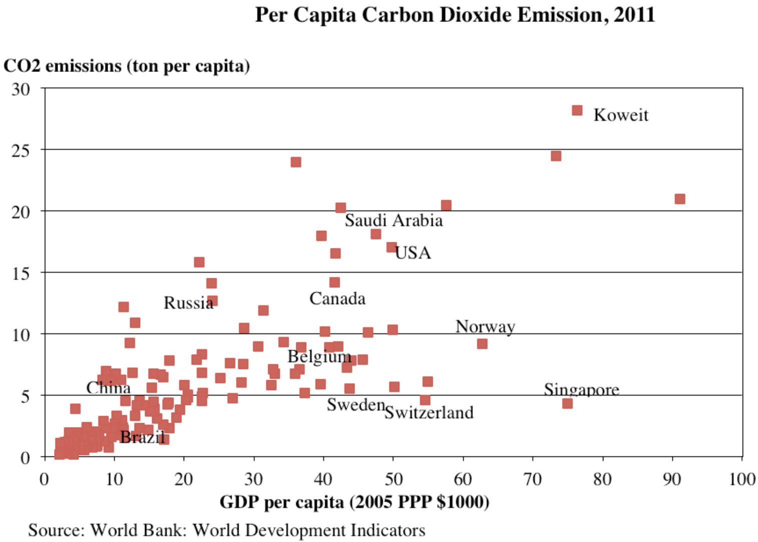
GDP and CO2 scatter plot



Data

- We have *Data*
    - The data are a matrix of observations y_i, x_i
    - y_i is the dependent variable
    - x_i is the independent variable
    - Each row of data corresponds to a unit i
- In this case, i is a country, y_i is $\frac{CO_{2i}}{Pop_i}$, x_i is $\frac{GDP_i}{Pop_i}$

How can we model this relationship?



Why do we need a model?

- To make sense of data, we need a statistical *model*

$$y_i = f(x_i) + \epsilon_i \tag{1}$$

$$\frac{CO_{2i}}{Pop_i} = f(\frac{GDP_i}{Pop_i}) + \epsilon_i \tag{2}$$

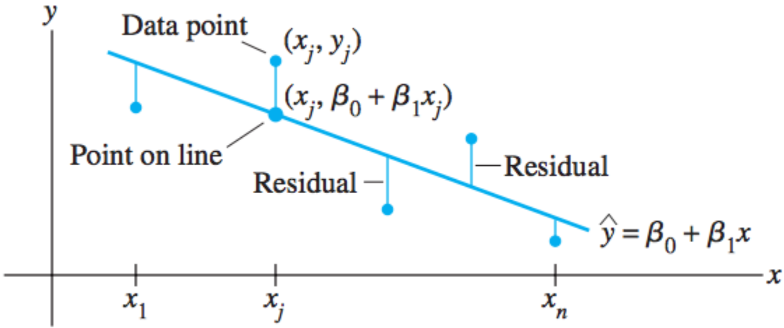
The linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3}$$

$$\frac{CO_{2i}}{Pop_i} = \beta_0 + \beta_1 \frac{GDP_i}{Pop_i} + \epsilon_i \tag{4}$$

- The linear model tells us that we think the relationship between CO_2 / Pop and GDP / Pop is *linear*
    - There is a constant amount of CO_2 / Pop emitted by all countries (β_0)
    - Each extra unit of GDP / Pop is associated with an additional β_1 units of $CO_2 / Capita$
- We estimate the parameters of the linear model via a tool called *linear regression* or *Ordinary Least Squares*

Linear regression, graphically



Causality

- When can we draw *causal* interpretations?
  - What is the effect of an increase in GPD/capita on CO2/capita, holding all else constant?
- This class will teach you what conditions must hold for a modeled relationship to identify causal effects, as opposed to just a statistical association
- At the end of class you will be able to answer: Does X cause Y holding all else constant?
  - I reject that X causes Y holding all else constant with a certain level of confidence
  - I cannot reject that X causes Y holding all else constant with a certain level of confidence
- Preview: functional form and omitted variables

Causality: Functional form

- What if the relationship between $\frac{CO_{2i}}{Pop_i}$ and $\frac{GDP_i}{Pop_i}$ is not linear?
- We can adjust the statistical model: for any $f(\cdot)$, we can specify

$$\frac{CO_{2i}}{Pop_i} = \beta_0 + \beta_1 f(\frac{GDP_i}{Pop_i}) + \epsilon_i \tag{5}$$

- What we need for causality: True model is linear in a *function* of x

Causality: Omitted variables

- Suppose something else also matters for $\frac{CO_2}{Pop}$
- What if we know about it?
- Suppose our dataset consists of $\frac{CO_{2i}}{Pop_i}$, $\frac{GDP_i}{Pop_i}$ but also some other variables $x_{2i}, x_{3i}, ..., x_{ki}$
- We can adjust the statistical model

$$\frac{CO_{2i}}{Pop_i} = \beta_0 + \beta_1 \frac{GDP_i}{Pop_i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + \epsilon_i \tag{6}$$

- But what if something else matters for $\frac{CO_2}{Pop}$, and we don't have data about it, or even know about it?

Preview: Different type of data

Four main types of data:

- Cross-Sectional Data
  - Each observation is a different individual/firm/country/etc. in the same time period
- Repeated Cross-Section
  - Each Observation is a different individual/firm/country/etc., but data cover multiple time periods
- Panel Data
  - Multiple observations of the same individuals/firms/countries/etc. at different points in time
- Time Series (*not covered in this course*)
  - Multiple time periods of one individual/firm/country/etc.

We will use different techniques to analyze different types of data, which is why the textbook is organized by types of data