# EEP/IAS 118 - Introductory Applied Econometrics, Lecture 1

Gregory Lane

June 2017

# Intro

- Attendance
- Course Overview (syllabus)
- Econometrics Intro
  - Material can be hard to grasp initially - but that's okay
- Assignments
  - Work on assignments early!
  - The class is compressed (and there is no GSI), so I will not be able available to help as much as I would like just before assignments are due
  - First problem set due next Monday (6/26)
  - First Quiz next Tuesday (6/27)

# What is Econometrics?

Econometrics is a tool used to accomplish several possible goals:

1. Establishing relationships between two variables $x$ and $y$
   - E.g. between smoking and lung cancer
   - **Causality:** We are most interested in how much of a change in $y$ is *caused* by a change in $x$. This can be hard: we want to say smoking *causes* cancer rather than people who smoke are more likely to die of cancer.
   - People who smoke might be poorer and live in less good environments which are also responsible for cancer

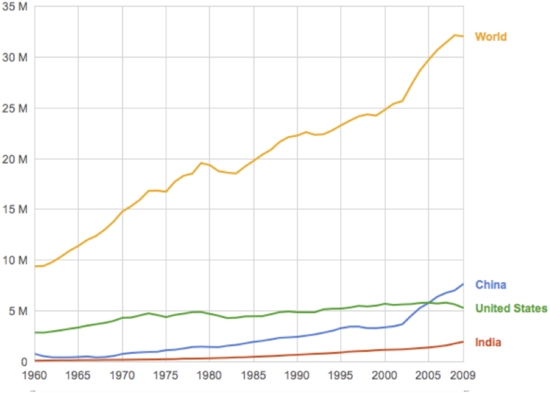
# What is Econometrics?

2. Evaluating a policy: E.g. mosquito nets distribution on malaria rates
    - We might also want to know does this have an effect on human capital - both long and short term
3. Testing a theory:
    - Instituting paid maternity leave increases the number of women in the labor force
    - Charter schools lead to higher success rates than public schools
    - Low income countries grow faster than high income countries

# Example: GDP and CO2 Emissions

As poorer countries become richer, how much will their growth contribute to this problem? This can help inform policy makers how much CO2 reduction needs to be achieved in order to reduce overall emissions.

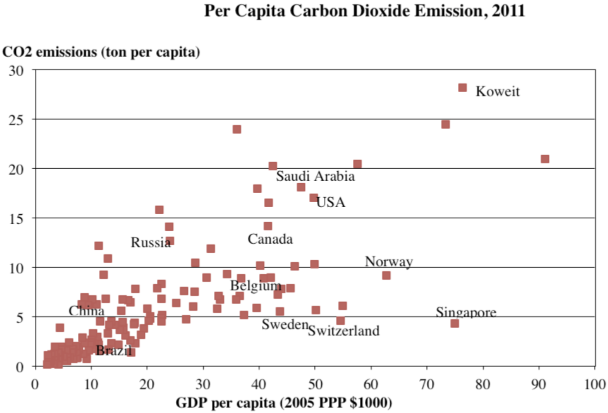
Total CO2 emission over time (M of kt = billions of tons)



Source: http://www.google.com/publicdata (World Bank Indicators)

# Example: GDP and CO2 Emissions

Plot GDP per capita and CO2 per capita together:



Source: World Bank: World Development Indicators

# Example: GDP and CO2 Emissions

**Goal:** We want to use this data to find the relationship between CO2 and GDP

- Need to create a *model* of this relationship
- A model is just some equation that relates the $x$ we are interested in to the $y$ we are interested in:

$$CO2/cap = f(GDP/cap)$$

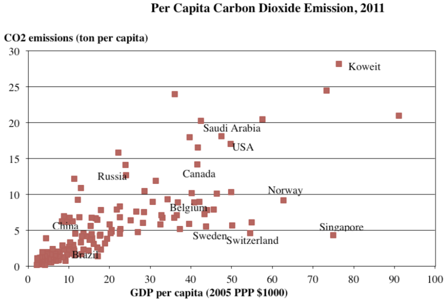
- In Econometrics, the workhorse model we use is the linear regression model:

$$\begin{aligned} CO2/cap &= f(GDP/cap) \\ &= \beta_0 + \beta_1(GDP/cap) \end{aligned}$$

# Example: GDP and CO2 Emissions

Fit our model, $CO2/cap = \beta_0 + \beta_1(GDP/cap)$, to the data

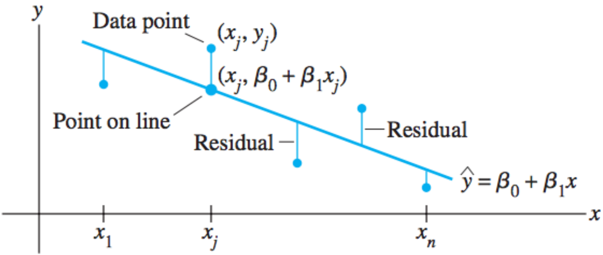
We are trying to draw a straight line through the data that best describes the relationship



Source: World Bank: World Development Indicators

Using the results of our model: "China's emission will increase by $y$% as it's GDP/cap increases by $x$%"

# Linear Regression Models: Overview



Key components of figure:

- Actual data: $x_i, y_i$ - observations of the two variables
- Line equation for $\hat{y} = \beta_0 + \beta_1 x$ - this is the predicted value of the outcome variable $y$
- Residuals ($\hat{u}_i$) - the difference between the predicted $\hat{y}_i$ and the actual observed $y_i$

# Altering the Model

- We have focused on relating one variable $x$ to one variable $y$. However, we can include many other factors that relate to the outcome in our model.
  - E.g. For CO2 emissions we might include production structure, climate, distance between population centers, etc.

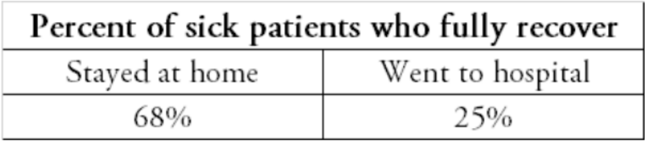
$$CO2/cap = \beta_0 + \beta_1(GDP/cap) + \beta_2 X_2 + ... + \beta_n X_n$$

- The model does *not* have to be linear.
  - Using a linear model is done mostly for convenience and ease of estimation
  - However, we shall see that the linear regression model can handle many types of relationships

# Problems with Causality

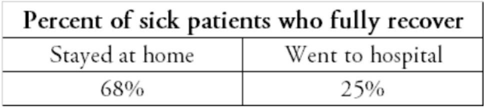
As we said, model is simply trying to describe a relationship between variables. However, we need to be careful.

A newspaper article states ”As you can see, health services here are so bad that going to a hospital is actually worse than staying at home. The following statistics demonstrate that you are better off staying away from hospitals”



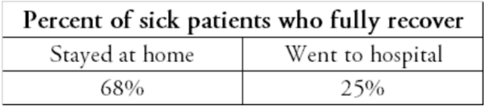
| Percent of sick patients who fully recover | |
|---|---|
| Stayed at home | Went to hospital |
| 68% | 25% |

A newspaper article states ”As you can see, health services here are so bad that going to a hospital is actually worse than staying at home. The following statistics demonstrate that you are better off staying away from hospitals”



| Percent of sick patients who fully recover | |
|---|---|
| Stayed at home | Went to hospital |
| 68% | 25% |

- What is the implied research question from this story?
- Do you agree with the news anchor’s conclusion? Why or why not?
- What are the components of the regression model you would use to analyze this question (if you had the data)?



| Percent of sick patients who fully recover | |
|---|---|
| Stayed at home | Went to hospital |
| 68% | 25% |

- What is the implied research question from this story? *What is the effect of going to the hospital on full recovery from an illness?*
- Do you agree with the news anchor's conclusion? *No, because the sample of people who go to the hospital is different from the sample that does not.*
- What are the components of the regression model you would use to analyze this question (if you had the data)?
  - *Dependent variable ($Y$) = Fully Recover*
  - *Explanatory variable of interest ($X_1$) = Went to hospital*
  - *Other explanatory variables ($X_1, X_2, ...$) = Age, Medical History, Severity of illness*

# Assigning Causality

- In this example, we do see a negative correlation between recovery and visiting the hospital
- So, what does newspaper article get wrong? There *is* a correlation!
    - The article falsely assigns *causality* to the relationship - this is the classic correlation $\neq$ causation
- The statistic is misleading (if improperly understood) because it omits other important variables associated with recovery from the model (age, medical history, severity of illness, etc.)
- **Key concept:** *Ceteris Paribus* ("All else equal") - we want to know the effect of going to the hospital on recovery *holding everything else constant*

# Data Types

We will be using several types of data throughout this course

1. **Cross-section:** We observe data ( $y$ and $x$es) for many units (households, individuals, firms) at a single point in time
    - Can observe correlations in this type of data
    - Very hard to establish casuality
2. **Time Series:** Have data for a country or household over time
    - We will deal with this data type near the end of the course

# Data Types

3. **Repeated Cross-section:** data from surveys across many points in time
    - However the observations in one year are not necessarily the same observations as before
4. **Panel:** Data on the *same* observations (households, individuals, firms) across many points in time
    - This is a useful type of data which we will see later in the course