Introduction to Data

Prottoy A. Akbar

Principles of Empirical Analysis (ECON-A3000)
Lecture 1

Breaking news!

Storks Deliver Babies ($p = 0.008$)

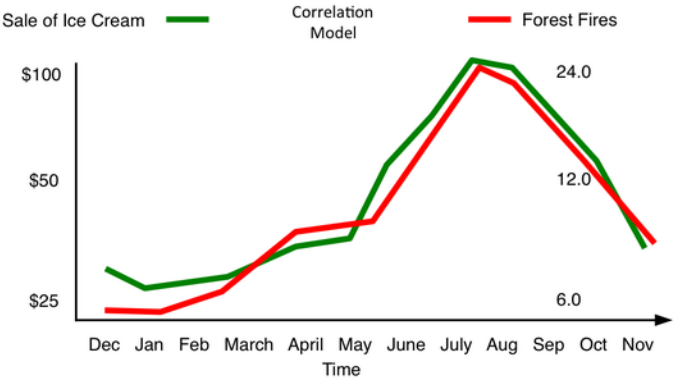
KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

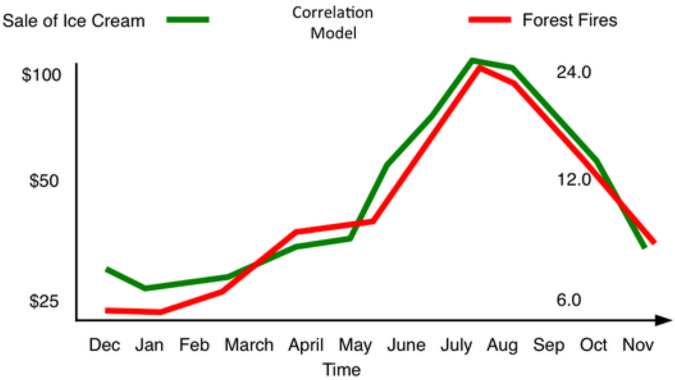
Source: Matthews, R. (2000), Storks Deliver Babies (p= 0.008). Teaching Statistics, 22: 36-38.

Ice creams cause forest fires?



Source: https://www.youtube.com/watch?v=VMUQSMFGBDo

Ice creams cause forest fires?

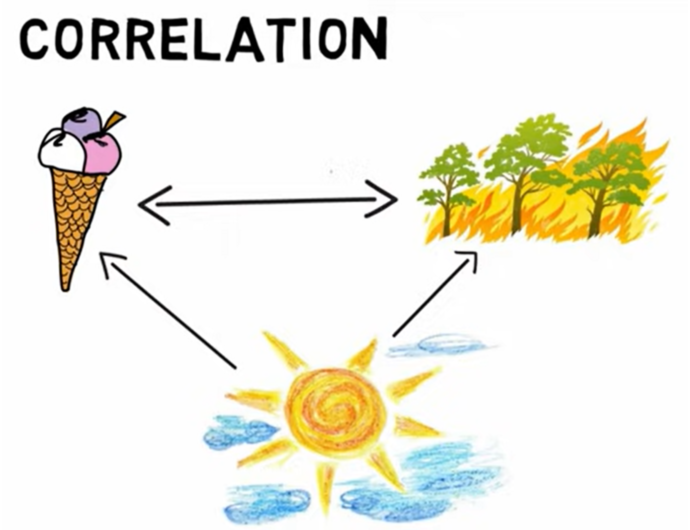


Source: https://www.youtube.com/watch?v=VMUQSMFGBDo

But also:

- obesity
- higher crime rates
- death by drowning

Ice creams cause forest fires?



Source: https://www.youtube.com/watch?v=VMUQSMFGBDo

This course

- This course is about ways to learn about the world through observation and experimentation:
  - moving beyond anecdotal evidence
  - using statistical tools to discern trends and patterns in data

This course

- This course is about ways to learn about the world through observation and experimentation:
    - moving beyond anecdotal evidence
    - using statistical tools to discern trends and patterns in data
- Aims to *complement* MS-A0503 (or equivalent course in probability and statistics)
    - It is not a math course
    - but: meaningful empirical inquiry is impossible without the math
    - we will learn a few concepts and some notation along the way, but most of the math will be covered at MS-A0503
    - we will focus on using the math to make sense of real-world data

This course

- This course is about ways to learn about the world through observation and experimentation:
  - moving beyond anecdotal evidence
  - using statistical tools to discern trends and patterns in data
- Aims to *complement* MS-A0503 (or equivalent course in probability and statistics)
  - It is not a math course
  - but: meaningful empirical inquiry is impossible without the math
  - we will learn a few concepts and some notation along the way, but most of the math will be covered at MS-A0503
  - we will focus on using the math to make sense of real-world data
- Here, we learn to ask questions such as
  - What data do I need to answer this question empirically?
  - How precise are my estimates? Do I have enough statistical power?
  - Does this particular correlation imply causation?
  - What are the identifying assumptions of this research design?

Why this course?

- Economics is increasingly empirical
  - empirical work dominates some subfields
  - important for all branches of economics

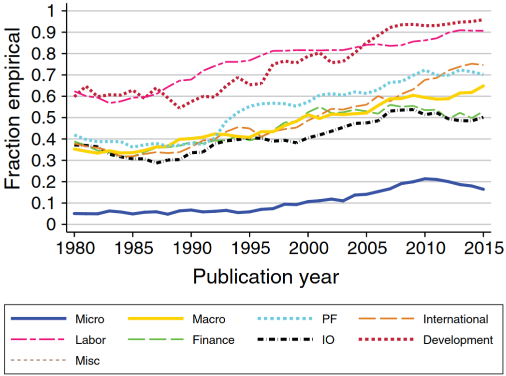


FIGURE 4. WEIGHTED FRACTION EMPIRICAL BY FIELD

Source: Angrist, J., Azoulay, P., Ellison, G., Hill R. and S. Lu 2017. Economic research evolves: Fields and styles. American Economic Review, Papers and Proceedings, 107, 5, 293-297.

Why this course?

- Economics is increasingly empirical
  - empirical work dominates some subfields
  - important for all branches of economics
- Theory and empirics are complements
  - theories need to be tested and quantified
  - empirical findings need to interpreted and generalized

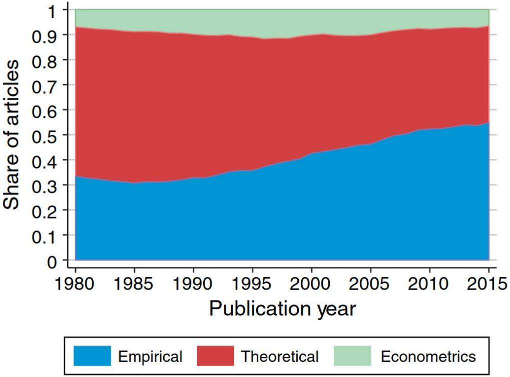
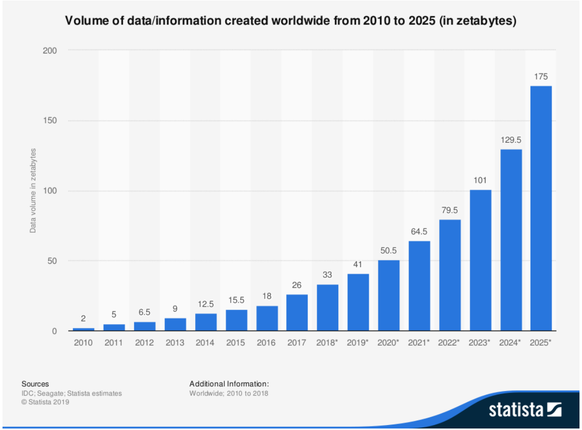


FIGURE 6. WEIGHTED PUBLICATIONS BY STYLE

Source: Angrist, J., Azoulay, P., Ellison, G., Hill R. and S. Lu 2017. Economic research evolves: Fields and styles. American Economic Review, Papers and Proceedings, 107, 5, 293-297.

Why this course?

- Economics is increasingly empirical
  - empirical work dominates some subfields
  - important for all branches of economics
- Theory and empirics are complements
  - theories need to be tested and quantified
  - empirical findings need to interpreted
- New opportunities are constantly emerging due to
  - more (digital) data becoming available
  - improvements in computing power
- But, old mistakes are still being made
  - more data is wonderful, but not a cure-all



Source: Statista.com. 1 zetabite = 1 billion terabytes = 1000^7 bytes.

Types of empirical research

- Three complementary approaches
  1. Descriptive: summarizing data, establishing facts
  2. Causal: how X *affects* Y?
  3. Prediction: how X *predicts* Y?

Types of empirical research

- Three complementary approaches
  1. Descriptive: summarizing data, establishing facts
  2. Causal: how X *affects* Y?
  3. Prediction: how X *predicts* Y?

- Example: ice cream consumption and forest fires
  - descriptive: strong correlation between the two
  - *not* causal: banning ice cream would probably not reduce forest fires
  - prediction: if all we observed was ice cream sales, we probably should use it for preparing for forest fires

Types of empirical research

- Three complementary approaches
  1. Descriptive: summarizing data, establishing facts
  2. Causal: how X *affects* Y?
  3. Prediction: how X *predicts* Y?
- Example: ice cream consumption and forest fires
  - descriptive: strong correlation between the two
  - *not* causal: banning ice cream would probably not reduce forest fires
  - prediction: if all we observed was ice cream sales, we probably should use it for preparing for forest fires
- This course is about descriptive and causal work
  - reflects the focus of most economics research
  - data science more focused on prediction

Course Plan

Course homepage

Today

- The power and sources of data
  - Data-driven decision making is in vogue
  - Example 1: Causes of cholera
  - Example 2: Smoking causes lung cancer?
  - Example 3: Congested vs slow cities
  - Types of data sources

- Describing data
  - mean, median, quantiles
  - variance and standard deviation

The power and sources of data

Example: Causes of cholera

Cholera in Victorian London

- Cholera arrived in London in 1831
  - "The combination of scary symptoms and fear of the unknown seized the public's imagination and chlolera was characterised as a foreign epidemic (it was commonly known as Asiatic cholera), which was 'invading' the nation."

Source: Science Museum

Cholera in Victorian London

- Cholera arrived in London in 1831
  - "The combination of scary symptoms and fear of the unknown seized the public's imagination and chlolera was characterised as a foreign epidemic (it was commonly known as Asiatic cholera), which was 'invading' the nation."
- Competing theories of cholera's causes
  - miasmas: particles in the air from decaying matter ("smell is disease")
  - germs: unknown germ transmitted by individuals ingesting water
- Both consistent with London's extremely bad sanitation conditions at the time

Source: Science Museum

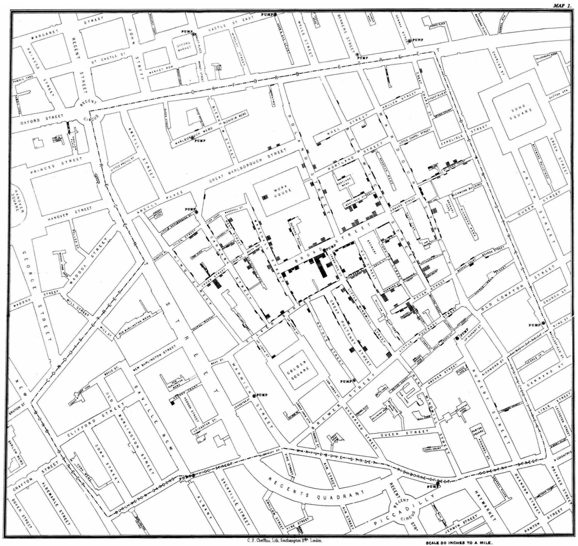
John Snow and the 1854 Broad Street Outbreak

- A particularly severe outbreak occurred in 1854 near Broad Street in Soho
  - 127 people died in three days
- John Snow identified the source as the public water pump on Broad Street and convinced authorities to disable it by removing its handle
  - initially: talking to local residents

John Snow memorial on Broadwick Street, Soho, London.

John Snow and the 1854 Broad Street Outbreak

- A particularly severe outbreak occurred in 1854 near Broad Street in Soho
  - 127 people died in three days
- John Snow identified the source as the public water pump on Broad Street and convinced authorities to disable it by removing its handle
  - initially: talking to local residents
  - later: map showing how cholera cases were clustered around this water pump
- This is just one example of how systematic data collection revolutionized medicine and public health



Original map by John Snow showing the clusters of cholera cases in the London epidemic of 1854. *Source:* Wikipedia.

The power and sources of data

Example: Smoking causes cancer?

Lung cancer in 1950s UK

- Lung cancer rates had increased six times in the last two decades
  - deaths exceeded deaths from tuberculosis for the first time.
  - even realizing this requires systematic data analysis

Lung cancer in 1950s UK

- Lung cancer rates had increased six times in the last two decades
  - deaths exceeded deaths from tuberculosis for the first time.
  - even realizing this requires systematic data analysis
- People commonly attributed it to rise in motor vehicles.
- How did we figure out it was cigarettes?

Doll and Hill smoking study 1950

First trial with twenty hospitals in north-west London.

- For each new cancer patient, nurses found - at random - another patient in the same hospital of the same sex and about the same age.
- Quiz both cancer patients and their counterparts about where they lived and worked, their lifestyle and diet, and their history of smoking.

Doll and Hill smoking study 1950

First trial with twenty hospitals in north-west London.

- For each new cancer patient, nurses found - at random - another patient in the same hospital of the same sex and about the same age.
- Quiz both cancer patients and their counterparts about where they lived and worked, their lifestyle and diet, and their history of smoking.
- Less than 2 years after trial began, Doll stopped smoking.

Doll and Hill smoking study 1950

First trial with twenty hospitals in north-west London.

- For each new cancer patient, nurses found - at random - another patient in the same hospital of the same sex and about the same age.
- Quiz both cancer patients and their counterparts about where they lived and worked, their lifestyle and diet, and their history of smoking.
- Less than 2 years after trial began, Doll stopped smoking.
- They discovered that heavy smoking made you **X** times more likely to get lung cancer!

Doll and Hill smoking study 1950

First trial with twenty hospitals in north-west London.

- For each new cancer patient, nurses found - at random - another patient in the same hospital of the same sex and about the same age.
- Quiz both cancer patients and their counterparts about where they lived and worked, their lifestyle and diet, and their history of smoking.
- Less than 2 years after trial began, Doll stopped smoking.
- They discovered that heavy smoking made you **X** times more likely to get lung cancer!

Doll and Hill smoking study 1950

First trial with twenty hospitals in north-west London.

- For each new cancer patient, nurses found - at random - another patient in the same hospital of the same sex and about the same age.
- Quiz both cancer patients and their counterparts about where they lived and worked, their lifestyle and diet, and their history of smoking.
- Less than 2 years after trial began, Doll stopped smoking.
- They discovered that heavy smoking made you **X** times more likely to get lung cancer!

Your responses

Doll and Hill smoking study 1950

First trial with twenty hospitals in north-west London.

- For each new cancer patient, nurses found - at random - another patient in the same hospital of the same sex and about the same age.
- Quiz both cancer patients and their counterparts about where they lived and worked, their lifestyle and diet, and their history of smoking.
- Less than 2 years after trial began, Doll stopped smoking.
- They discovered that heavy smoking made you **sixteen** times more likely to get lung cancer!

Your responses

Doll and Hill smoking study 1950

First trial with twenty hospitals in north-west London.

- For each new cancer patient, nurses found - at random - another patient in the same hospital of the same sex and about the same age.
- Quiz both cancer patients and their counterparts about where they lived and worked, their lifestyle and diet, and their history of smoking.
- Less than 2 years after trial began, Doll stopped smoking.
- They discovered that heavy smoking made you **sixteen** times more likely to get lung cancer!
- Very large effect! Good reason to be skeptical...

Your responses

Doll and Hill smoking study 1954

Longer-term and more ambitious trial: EVERY doctor in the UK (almost 60,000)

- were asked to complete a questionnaire about their health and smoking habits.

Doll and Hill smoking study 1954

Longer-term and more ambitious trial: EVERY doctor in the UK (almost 60,000)

- were asked to complete a questionnaire about their health and smoking habits.
- Doctors stay on the medical register, so they are easy to find and follow up on. They are also more capable of self-reporting and keeping track of their health and smoking.

Doll and Hill smoking study 1954

Longer-term and more ambitious trial: EVERY doctor in the UK (almost 60,000)

- were asked to complete a questionnaire about their health and smoking habits.
- Doctors stay on the medical register, so they are easy to find and follow up on. They are also more capable of self-reporting and keeping track of their health and smoking.
- $> 40,000$ doctors responded and 85% of the initial sample were smokers.

Doll and Hill smoking study 1954

Longer-term and more ambitious trial: EVERY doctor in the UK (almost 60,000)

- were asked to complete a questionnaire about their health and smoking habits.
- Doctors stay on the medical register, so they are easy to find and follow up on. They are also more capable of self-reporting and keeping track of their health and smoking.
- $> 40,000$ doctors responded and 85% of the initial sample were smokers.

One doctor buttonholed Hill at a London party. 'You're the chap who wants us to stop smoking,' he pointedly declared. 'Not at all,' replied Hill, ... 'I'm interested if you go on smoking to see how you die. I'm interested if you stop because I want to see how you die. So you choose for yourself, stop or go on... I shall score up your death anyway.'

- Hartford, Tim (2020). *How to Make the World Add Up.* The Bridge Street Press.

Doll and Hill smoking study 1954

Longer-term and more ambitious trial: EVERY doctor in the UK (almost 60,000)

- were asked to complete a questionnaire about their health and smoking habits.
- Doctors stay on the medical register, so they are easy to find and follow up on. They are also more capable of self-reporting and keeping track of their health and smoking.
- $> 40,000$ doctors responded and 85% of the initial sample were smokers.

One doctor buttonholed Hill at a London party. 'You're the chap who wants us to stop smoking,' he pointedly declared. 'Not at all,' replied Hill, ... 'I'm interested if you go on smoking to see how you die. I'm interested if you stop because I want to see how you die. So you choose for yourself, stop or go on... I shall score up your death anyway.'

- Hartford, Tim (2020). *How to Make the World Add Up.* The Bridge Street Press.

- Goal is, not to make prescriptions for, but to learn more systematically about the world.

Doll and Hill smoking study 1954

Longer-term and more ambitious trial: EVERY doctor in the UK (almost 60,000)

- were asked to complete a questionnaire about their health and smoking habits.
- Doctors stay on the medical register, so they are easy to find and follow up on. They are also more capable of self-reporting and keeping track of their health and smoking.
- $> 40,000$ doctors responded and 85% of the initial sample were smokers.

One doctor buttonholed Hill at a London party. 'You're the chap who wants us to stop smoking,' he pointedly declared. 'Not at all,' replied Hill, ... 'I'm interested if you go on smoking to see how you die. I'm interested if you stop because I want to see how you die. So you choose for yourself, stop or go on... I shall score up your death anyway.'

- Hartford, Tim (2020). *How to Make the World Add Up.* The Bridge Street Press.

- Goal is, not to make prescriptions for, but to learn more systematically about the world.
- But Hill's response raises a different concern with such experiments:
  As doctors learn about the health costs of smoking, some choose to continue smoking while others stop. Are the two groups truly comparable?

Tobacco Industry strikes back!

"They muddled the waters. They questioned the existing research; they funded research into other things they might persuade the media to get excited about, ... They manufactured doubt. A secret industry memo later reminded insiders"

- Hartford, Tim (2020). *How to Make the World Add Up.* The Bridge Street Press.

Doubt is our product since it is the best means of competing with the "body of fact" that exists in the mind of the general public. It is also the means of establishing a controversy. Within the business we recognize that a controversy exists. However, with the general public the consensus is that cigarettes are in some way harmful to the health. If we are successful in establishing a controversy at the public level, then there is an opportunity to put across the real facts about smoking

690010954

"Smoking and Health Proposal", Brown and Williamson internal memo, 1969

Tobacco Industry strikes back!

In the spring of 1965, a US Senate committee was pondering the life-and-death matter of whether to put a health warning on packets of cigarettes. An expert witness wasn't so sure about the scientific evidence, and so he turned to the topic of storks and babies. There was a positive correlation between the number of babies born and number of storks in the vicinity, he explained. That old story about babies being delivered by storks wasn't true, the expert went on; of course it wasn't. Correlation is not causation... Similarly, just because smoking was correlated with lung cancer did not mean - not for a moment - that smoking caused cancer.

- Hartford, Tim (2020). *How to Make the World Add Up.* The Bridge Street Press.

Tobacco Industry strikes back!

In the spring of 1965, a US Senate committee was pondering the life-and-death matter of whether to put a health warning on packets of cigarettes. An expert witness wasn't so sure about the scientific evidence, and so he turned to the topic of storks and babies. There was a positive correlation between the number of babies born and number of storks in the vicinity, he explained. That old story about babies being delivered by storks wasn't true, the expert went on; of course it wasn't. Correlation is not causation... Similarly, just because smoking was correlated with lung cancer did not mean - not for a moment - that smoking caused cancer.

- Hartford, Tim (2020). *How to Make the World Add Up.* The Bridge Street Press.

- The witness, Darrell Huff, had been paid by the tobacco lobby. He was the author of the 1954 best-seller *How to Lie with Statistics.*

Investigate boldly and interpret cautiously

- It's important to be cautious with causal interpretations.
- But it's also easy to be cynical of causal relationships
    - especially when you are unfamiliar with the data setting or the empirical tools available to us for causal analysis.
    - People often exploit the cautiousness of researchers to sow doubt.

The power and sources of data

Example: Congested vs slow cities

TRAVEL | The Bangladeshi Traffic Jam That Never Ends

The Bangladeshi Traffic Jam That Never Ends

By JODY ROSEN SEPT. 23, 2016

Of all the dysfunctions that plague the world's megacities, none may be more pernicious than bad (really, really bad) traffic. Sitting still in Dhaka, where bad design takes on epic proportions.

Akbar, Couture, Duranton, and Storeygard (2023)

- Use Google Maps to collect info on about 600 million trips in 1,358 cities on all continents
  - 97% of global urban population outside China
- Produce city-level speed and congestion indices, comparable across cities
- Investigate the determinants of urban travel speed and congestion and provide insightful decompositions
  - why some countries are fast, some are slow and some are congested?
- Provide a global open database on urban transportation
  - very little existing data on urban mobility around the world, especially in poor countries.

How to compare speeds across cities?

How to compare speeds across cities?

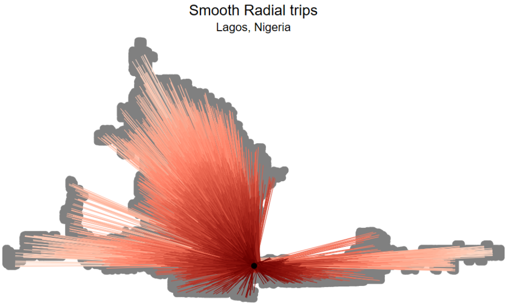
- Price index methodology
  - Each trip is a 'good'.
  - Speed is the (inverse) price of a trip in units of time.
  - Use a comparable basket of trips in each city.

How to compare speeds across cities?

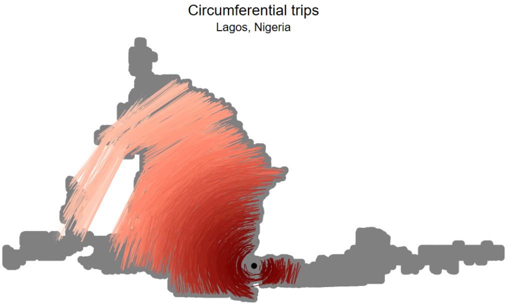
- Price index methodology
  - Each trip is a 'good'.
  - Speed is the (inverse) price of a trip in units of time.
  - Use a comparable basket of trips in each city.

- To obtain representative trips, we:
  - Design trips that resemble actual trips
  - Use different design strategies and verify they lead to similar results

Data: Illustration, trips for Lagos, Nigeria



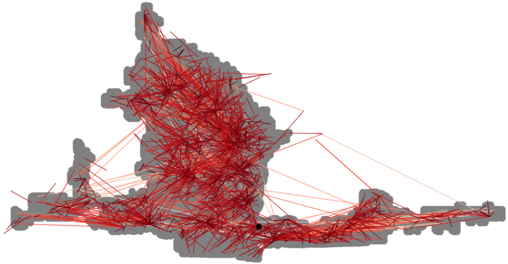
Radial trips



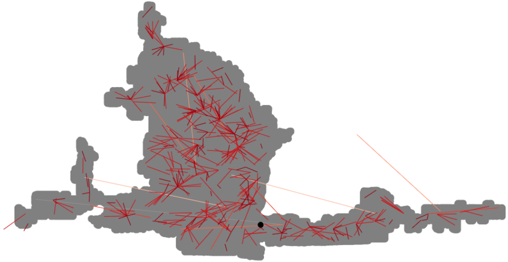
Circumferential trips

Akbar, Couture, Duranton and Storeygard (2023)

Data: Illustration, trips for Lagos, Nigeria



Trips to 'work'



Trips to school

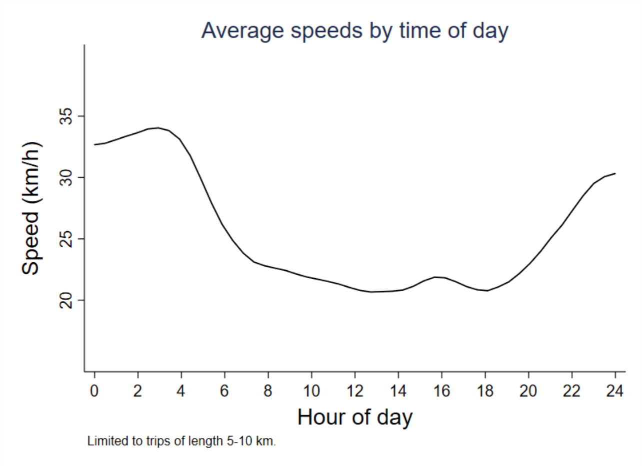
Akbar, Couture, Duranton and Storeygard (2023)

Combination of nearest and most popular according to Google

Sampling Trips

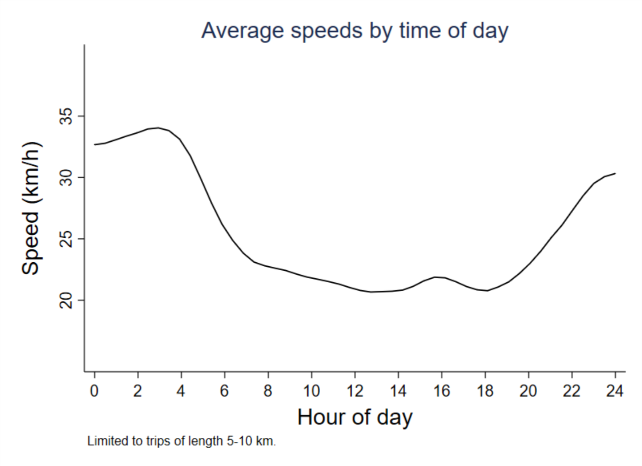
- About 20M trips in total and about 30 instances of each trip
  - at random times following a time/day distribution inspired by various travel surveys
- Simulated on Google Maps (website, GM)
  - "real time traffic" motor vehicle trip instances
  - between June and November 2019
  - leveraging on AWS computers
- For each trip instance and recommended GM route, we collect:
  - trip duration and length ($\Rightarrow$ speed)
  - duration in hypothetical state of no traffic ($\Rightarrow$ uncongested speed)

Mobility and congestion



Akbar, Couture,
Duranton and Storeygard (2023)

Mobility and congestion

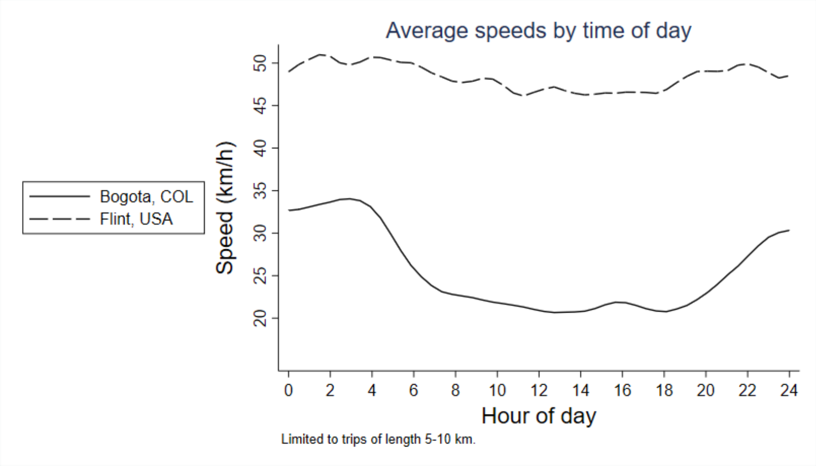


Akbar, Couture,
Duranton and Storeygard (2023)

Three mobility indices:

- Speed
- Uncongested speed
- Congestion

Mobility vs congestion

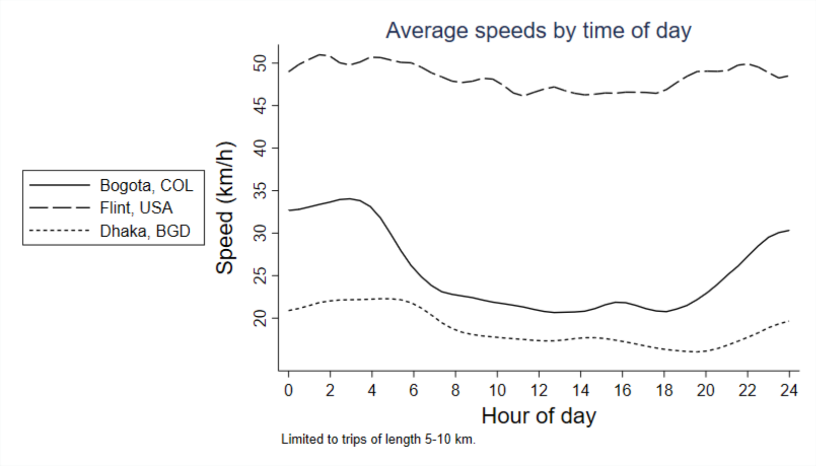


Akbar, Couture, Duranton and Storeygard (2023)

Three mobility indices:

- Speed
- Uncongested speed
- Congestion

Mobility vs congestion



Akbar, Couture, Duranton and Storeygard (2023)

Three mobility indices:

- Speed
- Uncongested speed
- Congestion

What do we expect a slow city to look like?

- Large cities?
- In wealthy countries with high car ownership?
- Old cities with narrow roads that are easily congestible?

Results: Fastest, slowest and most congested cities

	Fastest			Slowest			Most Congested		
Rank	City	Country	Index	City	Country	Index	City	Country	Index
1	Flint	United States	.47	Dhaka	Bangladesh	-.63	Bogotá	Colombia	.21
2	Greensboro	United States	.43	Lagos	Nigeria	-.58	Krasnodar	Russia	.20
3	Little Rock	United States	.43	Manila	Philippines	-.53	Moscow	Russia	.18
4	Wichita	United States	.42	Ikorodu	Nigeria	-.53	Bucharest	Romania	.18
5	Huntsville	United States	.41	Kolkata	India	-.51	Ulaanbaatar	Mongolia	.18
6	Lancaster-Palmdale	United States	.41	Bhiwandi	India	-.51	Manila	Philippines	.17
7	Victorville	United States	.40	Mumbai	India	-.45	Bangkok	Thailand	.17
8	Ogden	United States	.40	Phnom Penh	Cambodia	-.44	Bangalore	India	.17
9	Lansing	United States	.40	Chittagong	Bangladesh	-.43	Vladivostok	Russia	.15
10	Knoxville	United States	.38	Bangalore	India	-.43	Mexico City	Mexico	.15
11	Visalia	United States	.38	Dar es Salaam	Tanzania	-.43	London	United Kingdom	.15
12	Tulsa	United States	.38	Kumasi	Ghana	-.43	Lagos	Nigeria	.15
13	Khamis Mushayt	Saudi Arabia	.38	Jakarta	Indonesia	-.42	Mumbai	India	.14
14	Shreveport	United States	.37	Aba	Nigeria	-.42	Yekaterinburg	Russia	.14
15	Winston-Salem	United States	.37	Bihar Sharif	India	-.42	Guatemala City	Guatemala	.14
16	Port St. Lucie	United States	.37	Arrah	India	-.42	New York	United States	.14
17	Youngstown	United States	.37	Bacoor	Philippines	-.41	Delhi	India	.13
18	Toledo	United States	.36	Mymensingh	Bangladesh	-.41	Sochi	Russia	.13
19	Fayetteville-Springdale	United States	.36	Patna	India	-.41	Panama City	Panama	.13
20	Rockford	United States	.36	Lima	Peru	-.41	Nairobi	Kenya	.13

Akbar, Couture, Duranton and Storeygard (2023)

What do we expect a slow city to look like?

- Large cities?
- In wealthy countries with high car ownership?
- Old cities with narrow roads that are easily congestible?

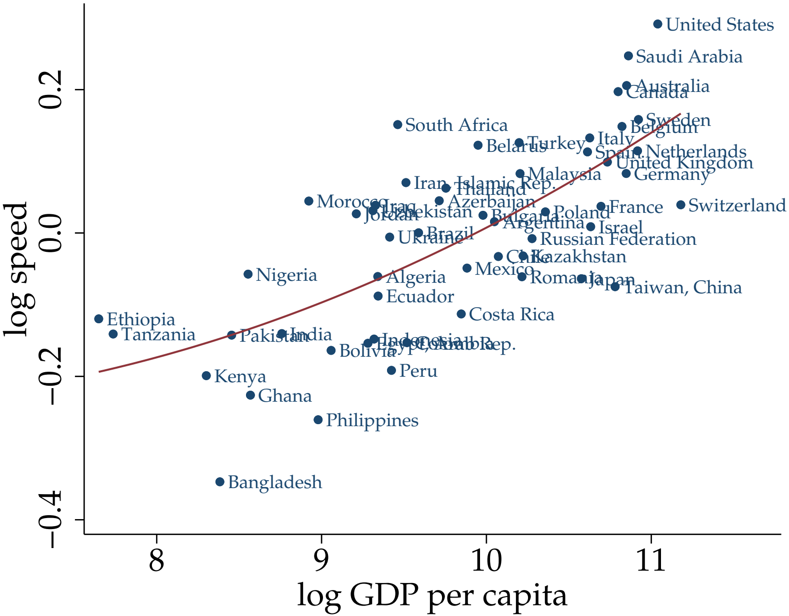
What do we expect a slow city to look like?

- Large cities?
- In wealthy countries with high car ownership?
- Old cities with narrow roads that are easily congestible?

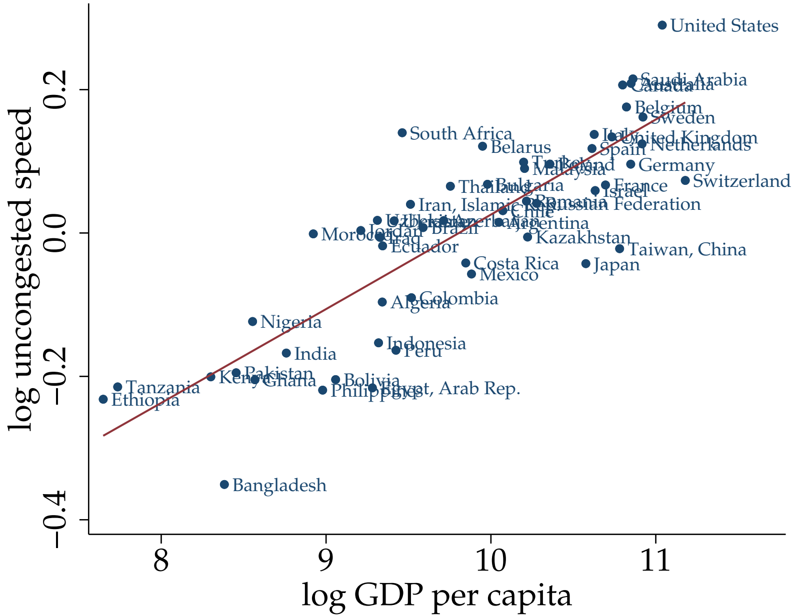
We find the biggest predictor of driving speeds to be:

Country wealth

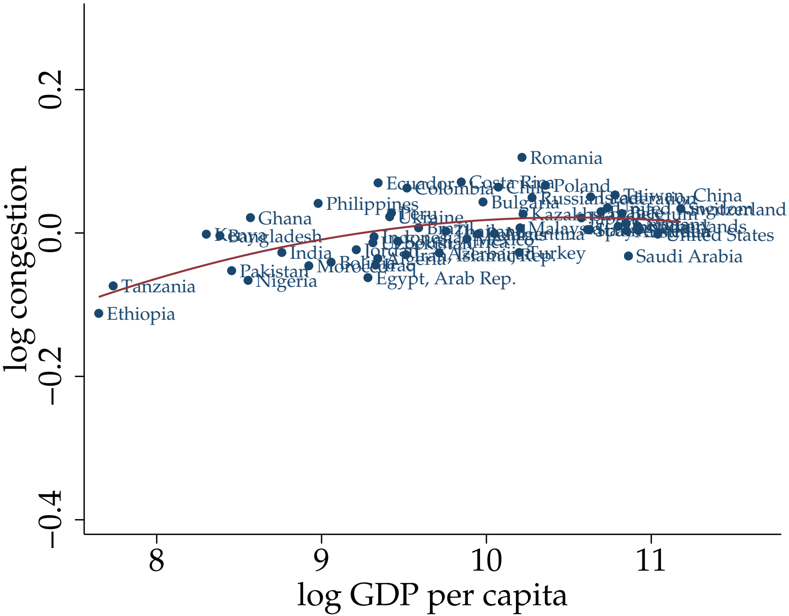
Speed vs. GDP pc, country



Uncongested speed vs. GDP pc, country



Congestion vs. GDP pc,country



Congestion is not that important empirically!

Congestion has been the primary focus of transportation economists and engineers for decades

- theoretical focus based on anecdotal rather than empirical evidence
- has dominated policy-making: Dhaka has tried limiting cars on the road, banning slower vehicles like bicycle rickshaws, etc.

Congestion is not that important empirically!

Congestion has been the primary focus of transportation economists and engineers for decades

- theoretical focus based on anecdotal rather than empirical evidence
- has dominated policy-making: Dhaka has tried limiting cars on the road, banning slower vehicles like bicycle rickshaws, etc.

Is this paper making a causal claim?

Congestion is not that important empirically!

Congestion has been the primary focus of transportation economists and engineers for decades

- theoretical focus based on anecdotal rather than empirical evidence
- has dominated policy-making: Dhaka has tried limiting cars on the road, banning slower vehicles like bicycle rickshaws, etc.

Is this paper making a causal claim?

- e.g. as cities like Dhaka grow economically, will they get faster?

Congestion is not that important empirically!

Congestion has been the primary focus of transportation economists and engineers for decades

- theoretical focus based on anecdotal rather than empirical evidence
- has dominated policy-making: Dhaka has tried limiting cars on the road, banning slower vehicles like bicycle rickshaws, etc.

Is this paper making a causal claim?

- e.g. as cities like Dhaka grow economically, will they get faster?
- Not really causal. Many things differ between the cities besides income per capita.

Congestion is not that important empirically!

Congestion has been the primary focus of transportation economists and engineers for decades

- theoretical focus based on anecdotal rather than empirical evidence
- has dominated policy-making: Dhaka has tried limiting cars on the road, banning slower vehicles like bicycle rickshaws, etc.

Is this paper making a causal claim?

- e.g. as cities like Dhaka grow economically, will they get faster?
- Not really causal. Many things differ between the cities besides income per capita.
- Yet we can learn new things about the world.
  - In fact, we need to learn to first describe the relationships we observe in data before we can even start to make causal claims about them.
  - That's where we will start.

Data sources

Data sources

- National statistical offices and alike
  - permanent, standardized surveys
    - e.g. census, labor force surveys
  - administrative register data
    - e.g. tax register, population register

Census enumerators in the 19th century UK (top) and 2020 US (bottom).

Data sources

- National statistical offices and alike
  - permanent, standardized surveys
    - e.g. census, labor force surveys
  - administrative register data
    - e.g. tax register, population register
- Private proprietary data
  - e.g. cell phone locations, grocery store chains
  - sometimes shared with researchers

The New York Times Magazine

How Companies Learn Your Secrets

570

Antonio Bolfo/Reportage for The New York Times

By Charles Duhigg

Feb. 16, 2012

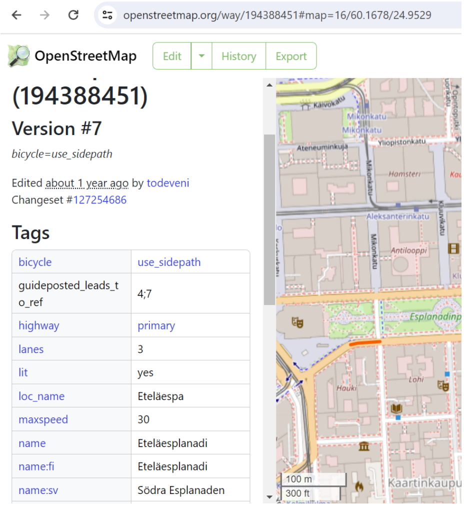
Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that? "

Pole has a master's degree in statistics and another in economics, and has been obsessed with the intersection of data and human behavior most of his life. His parents were teachers in North

Source: New York Times Magazine, 16 Feb 2012.

Data sources

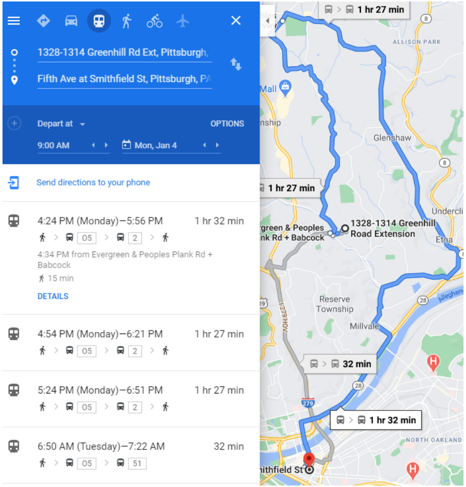
- National statistical offices and alike
  - permanent, standardized surveys
    - e.g. census, labor force surveys
  - administrative register data
    - e.g. tax register, population register
- Private proprietary data
  - e.g. cell phone locations, grocery store chains
  - sometimes shared with researchers
- Crowd-sourced data e.g. Wikipedia, Open Street Map, General Transit Feed Specification, ...



Source: OpenStreetMap

More data sources

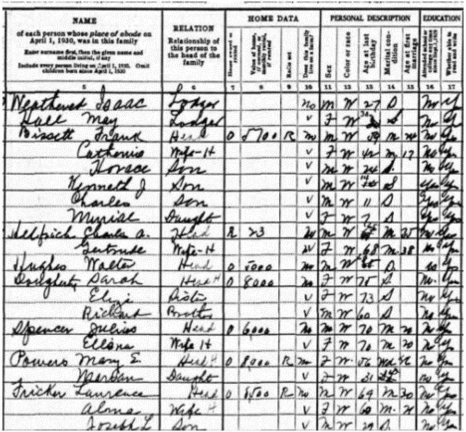
- Publicly available data that you already use e.g. navigation apps, online reviews, social media, property sale/rent ads,



Source: Google Maps

More data sources

- Publicly available data that you already use e.g. navigation apps, online reviews, social media, property sale/rent ads,
- Collect your own data e.g. surveys, archival records (digitized or need to be)
  - maybe others have already done this e.g. replication packages of publications



Source: 1930 Census

More data sources

- Publicly available data that you already use e.g. navigation apps, online reviews, social media, property sale/rent ads,
- Collect your own data e.g. surveys, archival records (digitized or need to be)
  - maybe others have already done this e.g. replication packages of publications
- Field and lab experiments
  - if you cannot identify an appropriate research setting in the real world, simulate it!

Photo source: Georgia State Experimental Economics Laboratory

Lab experiments

Register for Helsinki Labbet

Helsinki Labbet is a multidisciplinary laboratory for experimental studies.
By participating in the studies, you will contribute to scientific research and earn money.
Register by clicking **here** or scanning the QR code.

https://www.helsinkilabbet.fi/

Descriptive statistics

Income distribution

- Aim: learning to characterize distributions
- Example: income distribution
  - the learning objectives could be fulfilled with any distribution
  - but this one is particularly central to much research and policy debate
- today: basics using Statistics Finland's teaching data

Source: The Economist, 28 Nov 2019

Statistics Finland's teaching data

- We use Statistics Finland's teaching data for this lecture and some your exercises
  - random sample of the old Finnish Linked Employer-Employee dataset (FLEED)
    - now under the name FOLK for research purposes (taika.stat.fi)
  - lots of information about all working age residents living in Finland and their employers (data description here → Variable description)

Statistics Finland's teaching data

- We use Statistics Finland's teaching data for this lecture and some your exercises
  - random sample of the old Finnish Linked Employer-Employee dataset (FLEED)
    - now under the name FOLK for research purposes (taika.stat.fi)
  - lots of information about all working age residents living in Finland and their employers (data description here → Variable description)
- We will use annual earned income for this analysis
  - Statistics Finland's metadata: "Earned income is the **sum of earned and entrepreneurial income** received by households and income recipients during the year. The earned income concept of the income distribution statistics includes income items **taxed in taxation** both as **earned and capital income**."
  - initial source: Finland's Tax Authority

Statistics Finland's teaching data

- We use Statistics Finland's teaching data for this lecture and some your exercises
  - random sample of the old Finnish Linked Employer-Employee dataset (FLEED)
    - now under the name FOLK for research purposes (taika.stat.fi)
  - lots of information about all working age residents living in Finland and their employers (data description here → Variable description)
- We will use annual earned income for this analysis
  - Statistics Finland's metadata: "Earned income is the **sum of earned and entrepreneurial income** received by households and income recipients during the year. The earned income concept of the income distribution statistics includes income items **taxed in taxation** both as **earned and capital income**."
  - initial source: Finland's Tax Authority
- In teaching data, all income is
  1. rounded to the nearest 1,000 euros
  2. top-coded at 100,000

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?
  - first: let's look at the data

	vuosi	shtun	sukup	syntyv	svatva
83069	15	1	2	1987	21000
83070	15	2	2	1945	18000
83071	15	4	2	1993	7000
83072	15	6	2	1983	16000
83073	15	7	2	1952	.
83074	15	8	2	1947	30000
83075	15	9	1	1950	21000
83076	15	10	2	1994	2000
83077	15	11	2	1949	10000
83078	15	12	2	1957	8000
83079	15	14	1	1946	35000
83080	15	15	1	1940	17000
83081	15	16	1	1957	34000
83082	15	18	1	1965	17000
83083	15	19	1	1979	.
83084	15	20	1	1957	40000
83085	15	21	1	1949	16000
83086	15	22	1	1994	1000
83087	15	23	1	1947	14000
83088	15	24	2	1968	29000
83089	15	26	2	1995	0
83090	15	28	2	1964	18000
83091	15	29	2	1962	52000
83092	15	30	2	1961	12000
83093	15	31	1	1977	26000
83094	15	32	2	1945	28000
83095	15	33	1	1992	1000
83096	15	35	2	1976	21000
83097	15	36	1	1990	2000

Vars: 5 of 18 Order: Dataset   Obs: 6,244 of 89,312

Source: FLEED teaching data

```
browse shtun vuosi sukup syntyv svatva if vuosi==15
```

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?
  - first: let's look at the data
  - second: let's clean it a little bit

```
rename shtun id
gen year=1995+vuosi
gen woman=(sukup==2)
replace woman=.  if sukup==.
gen age=year-syntyv
rename svatva earn
keep if year==2010
order id year earn age woman
```

	id	year	earn	age	woman
83069	1	2010	21000	23	1
83070	2	2010	18000	65	1
83071	4	2010	7000	17	1
83072	6	2010	16000	27	1
83073	7	2010	.	58	1
83074	8	2010	30000	63	1
83075	9	2010	21000	60	0
83076	10	2010	2000	16	1
83077	11	2010	10000	61	1
83078	12	2010	8000	53	1
83079	14	2010	35000	64	0
83080	15	2010	17000	70	0
83081	16	2010	34000	53	0
83082	18	2010	17000	45	0
83083	19	2010	.	31	0
83084	20	2010	40000	53	0
83085	21	2010	16000	61	0
83086	22	2010	1000	16	0
83087	23	2010	14000	63	0
83088	24	2010	29000	42	1
83089	26	2010	0	15	1
83090	28	2010	18000	46	1
83091	29	2010	52000	48	1
83092	30	2010	12000	49	1
83093	31	2010	26000	33	0
83094	32	2010	28000	65	1
83095	33	2010	1000	18	0
83096	35	2010	21000	34	1
83097	36	2010	2000	20	0

Vars: 5 of 21 Order: Dataset   Obs: 6,244 of 89,312

Source: FLEED teaching data
`browse id year earn age woman`

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?
  - first: let's look at the data
  - second: let's clean it a little bit

```
rename shtun id
gen year=1995+vuosi
gen woman=(sukup==2)
replace woman=.  if sukup==.
gen age=year-syntyv
rename svatva earn
keep if year==2010
order id year earn age woman
```

  - still: 5,973 is an awful lot of numbers...
- We need to find ways to summarize the data in an informative, but parsimonious manner

	id	year	earn	age	woman
83069	1	2010	21000	23	1
83070	2	2010	18000	65	1
83071	4	2010	7000	17	1
83072	6	2010	16000	27	1
83073	7	2010	.	58	1
83074	8	2010	30000	63	1
83075	9	2010	21000	60	0
83076	10	2010	2000	16	1
83077	11	2010	10000	61	1
83078	12	2010	8000	53	1
83079	14	2010	35000	64	0
83080	15	2010	17000	70	0
83081	16	2010	34000	53	0
83082	18	2010	17000	45	0
83083	19	2010	.	31	0
83084	20	2010	40000	53	0
83085	21	2010	16000	61	0
83086	22	2010	1000	16	0
83087	23	2010	14000	63	0
83088	24	2010	29000	42	1
83089	26	2010	0	15	1
83090	28	2010	18000	46	1
83091	29	2010	52000	48	1
83092	30	2010	12000	49	1
83093	31	2010	26000	33	0
83094	32	2010	28000	65	1
83095	33	2010	1000	18	0
83096	35	2010	21000	34	1
83097	36	2010	2000	20	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Vars: 5 of 21 Order: Dataset   Obs: 6,244 of 89,312

Source: FLEED teaching data
`browse id year earn age woman`

Descriptive statistics

- **Descriptive statistics:** ways of summarizing information to make data understandable
  - objective: reduce the amount of numbers as much as possible while losing as little information as possible

Descriptive statistics

- **Descriptive statistics:** ways of summarizing information to make data understandable
  - objective: reduce the amount of numbers as much as possible while losing as little information as possible
- Let's start with Stata's summarize command

```
summarize earn, detail
```

(Stata also allows shortened format e.g. `sum earn, d`)

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Descriptive statistics

- **Descriptive statistics:** ways of summarizing information to make data understandable
  - objective: reduce the amount of numbers as much as possible while losing as little information as possible
- Let's start with Stata's summarize command

```
summarize earn, detail
```

(Stata also allows shortened format e.g. `sum earn, d`)

- It gives us the key descriptive statistics:
  - sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  - single number measures of variation
  - selected quantiles

```
                            earn
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%         1000              0
10%         3000              0       Obs               5,973
25%        10000              0       Sum of Wgt.       5,973

50%        21000                      Mean           23296.67
                        Largest       Std. Dev.      17163.61
75%        33000         100000
90%        45000         100000       Variance       2.95e+08
95%        55000         100000       Skewness       1.006775
99%        78000         100000       Kurtosis       4.340098
```

Source: FLEED teaching data

Measures of variation

- Variance:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Standard deviation:

$$SD(x) = \sqrt{Var(x)}$$

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Measures of variation

- Variance:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Standard deviation:

$$SD(x) = \sqrt{Var(x)}$$

- Say, we knew how much individuals spend on groceries. How would the variance on grocery expenditures compare to the variance on incomes?

Measures of variation

- Variance:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Standard deviation:

$$SD(x) = \sqrt{Var(x)}$$

- Say, we knew how much individuals spend on groceries. How would the variance on grocery expenditures compare to the variance on incomes?

Your responses

Measures of variation

- To be able to compare across variables, sometimes we normalize the standard deviation with mean. This is called the coefficient of variation. In this example:

$$CV(x) = \frac{SD(x)}{\bar{x}} = \frac{17,164}{23,297} = .74$$

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Summary

- This course is about doing and reading empirical research
  - complements introductory statistics
  - aim is to learn to ask the right questions and to build intuition
  - these are critical skills for a modern economist
- Today:
  - Introduction to empirical analysis
  - How to describe data variables

Summary

- This course is about doing and reading empirical research
  - complements introductory statistics
  - aim is to learn to ask the right questions and to build intuition
  - these are critical skills for a modern economist
- Today:
  - Introduction to empirical analysis
  - How to describe data variables
- Next lecture: Samples and distributions
  - **Pre-class assignment 1 due** 15 minutes before next lecture
  - Hold on to your name tags/placards
- Homework 1 **due Jan 17**
  - can already install Stata to get started
  - but wait till after next lecture to complete