Case Study Session 3: Scientifically Analysing Data

Dr. L. Sanna Stephan

2023

RUG Groningen

Learning objectives

You will learn...

- ... what it entails for us as economists that our data is "getting bigger"
- ... why prediction is important in policy and why machine learning and big data make this a powerful tool
- ... how the trade-off simplicity and complexity is illustrated as a curve-fitting problem

Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?

Introduction

- Authors: Linnet Taylor, Ralph Schroeder, & Eric Meyer (2014)
- Focus:
  - How Big Data can be used?
  - How Big Data may contribute to/ change the field?
- Goal: evaluate the challenges and rewards of using Big Data in economics

Background

- Working definition of Big Data: a change in the **scale and scope** of sources and the tools used to manipulate these sources
- The field of economics has been fairly slow to adopt Big Data.
- Many economists are sceptical of Big Data and prefer to stick to the existing methodology.

Advantages and drawbacks of Big Data

Advantages:

1. Data frequently available in real-time → 'nowcasting'
2. Large datasets make analysis more powerful due to
    - ...high sample size
    - ...high level of detail
3. Enables the observation of variables that had previously been difficult to observe, such as
    - ...social networks
    - ...spatial data

Drawbacks:

1. Unstructured
2. Measurement of variables not explicitly designed for research
    - Data is 'unclean'

Economists and Big Data

Economists suited to use Big Data!

- Economists have technical skills (e.g. statistics and coding knowledge)
- Big Data can tackle estimation challenges/ improve econometrics
- Many big data sets come from economic activity:
  - Financial transactions
  - Loyalty card data
  - Social network data
  - etc.

Interviews

- economists and data scientists that use Big Data express opinion
  1. For what purposes are Big Data used?
  2. What type of knowledge is the use of this data contributing to?

Defining Big Data

Big Data is a term everyone uses yet nobody can define.

- The number of observations considered 'large' depends on the field
- New methods for data analysis become crucial in working with Big Data:
  - 'It [Big Data] starts when you can't use Stata, I think'
  - Specific skills are needed

- Using Big Data is like 'turning up the microscope'
- Increased involvement of industry in academia (e.g. granular customer data)
- Potential mutual benefit

Rationales for Big Data adoption

- Big Data (i.a. from web) as alternative to conventional data
- 'Big Data economics' is emerging as a discipline outside conventional economics:
  - Some computer scientists apply their field-specific techniques to analyse economic phenomena
  - Example: the relationship between Twitter and the stock market

- Some economists use the real-time dimension of Big Data
- But not all: some draw samples from a Big Dataset and apply conventional statistical approaches
    - '[using sub-sample] is just as good as using the data itself.'
    - can re-sample

- Those that adopt Big Data utilise new technical approaches
- Those include:
  - Web scraping
  - Application: gathering online prices in Argentina to create a more transparent inflation measure
  - Semi-cleaned on-demand data
  - Application: Google Analytics to proxy demand for goods

Interpretation challenges

- Rethinking statistical significance
  - 'when you have a billion observations, everything's significant'
- Is Big Data essentially descriptive? Or is it possible to establish causality?
- Sharpening modelling techniques potentially more important than data size
- Alternative view: theory loses importance if everything can be measured

- Economists making predictions of macroeconomic indicators: no longer necessary?
    - Experiment currently underway to see whether Twitter outperforms economists in predictions
- Reevaluation of data mining: previously frowned upon, now gaining credibility
    - Data mining is now used to 'search for the right questions'
- A broader definition of 'good methodology' in economic research is needed

Access challenges

- Big Data is frequently owned by corporations
- Potentially increased divide between senior and junior researchers
    - That is: only senior researchers get the 'best' data
- Non-disclosure agreements limit replication
- Ethical issues in using detailed information on health, employment, behaviour for academic research
- Some economists get non-academic jobs to acquire 'privileged' data

Taylor, Schroeder, & Meyer (2014): conclusion

Big Data

- ... requires new methods because of its size and (often unstructured) nature,
- ... has major disruptive potential in economics:
  - more data/possibilities → new economic theory,
  - (some) econometric methods/software may become redundant,
  - AI/computer scientists may beat economist in forecasting,
- ... provides new opportunities for sampling and modelling precision,

Taylor, Schroeder, & Meyer (2014): conclusion

Big Data

- ... is often owned by companies.
  - ☠ **threat:** replicability limited,
  - ☠ **threat:** senior researchers may have privileged access,
  - ✓ **benefit:** increase cooperation between industry and academia,

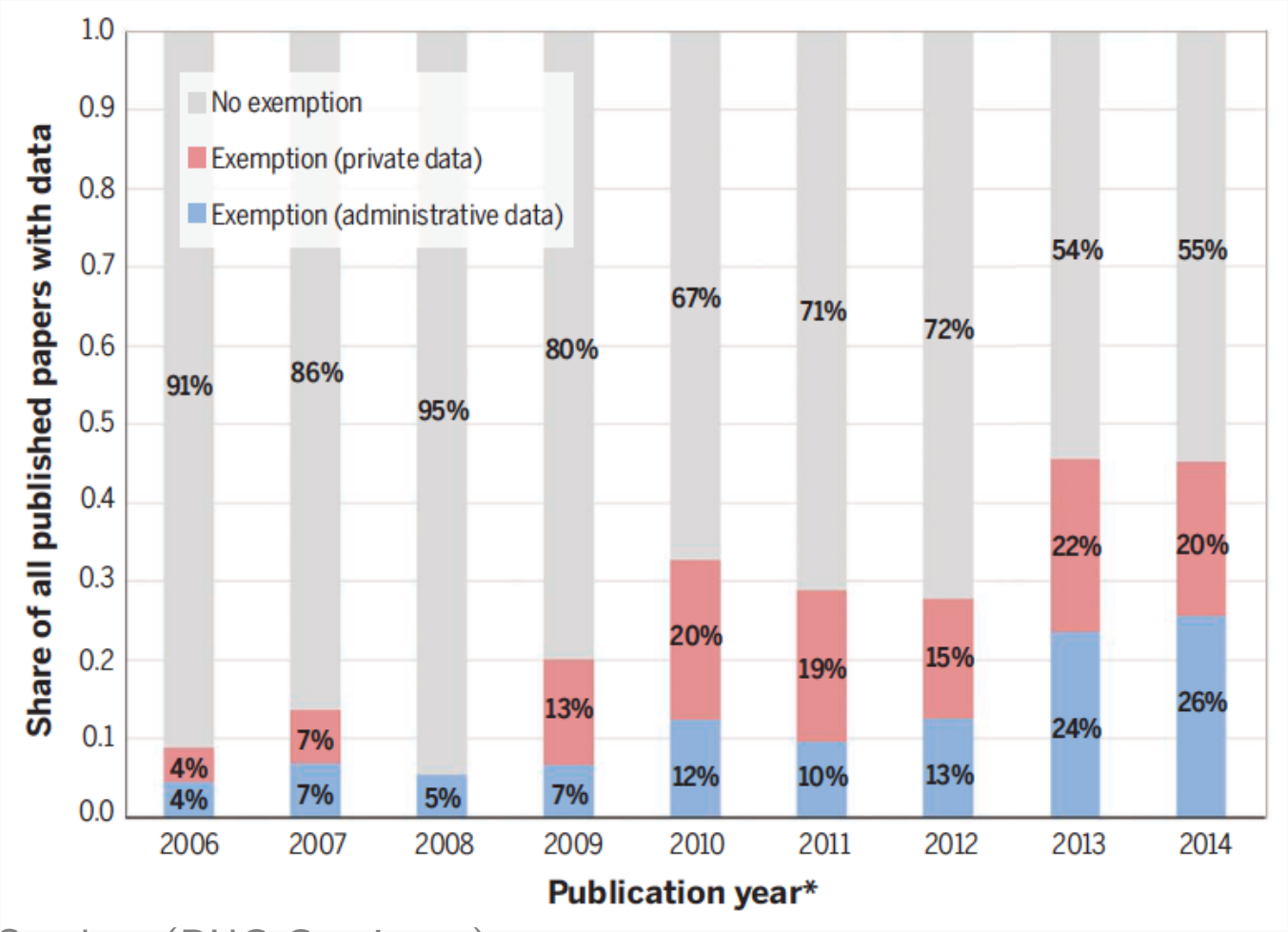
Economics in the age of Big Data

Introduction

- Authors: Liran Einav and Jonathan Levin (2014)
- Focus: Explore how Big Data affects economic research

Background

more non-publicly available data is used in economics (Journals require data to be published, or else authors must seek exemption)



1. Source of Big Data:
   - Administrative data
   - Private sector data
2. Statistical methods and role of theory

Administrative data

- Advantages over survey data:
  - Fewer missing observations
  - Large sample size
  - Many time periods
  - No/little sample selection
- Application: regional disparities in economic mobility in the United States

Administrative data: economic mobility in the US

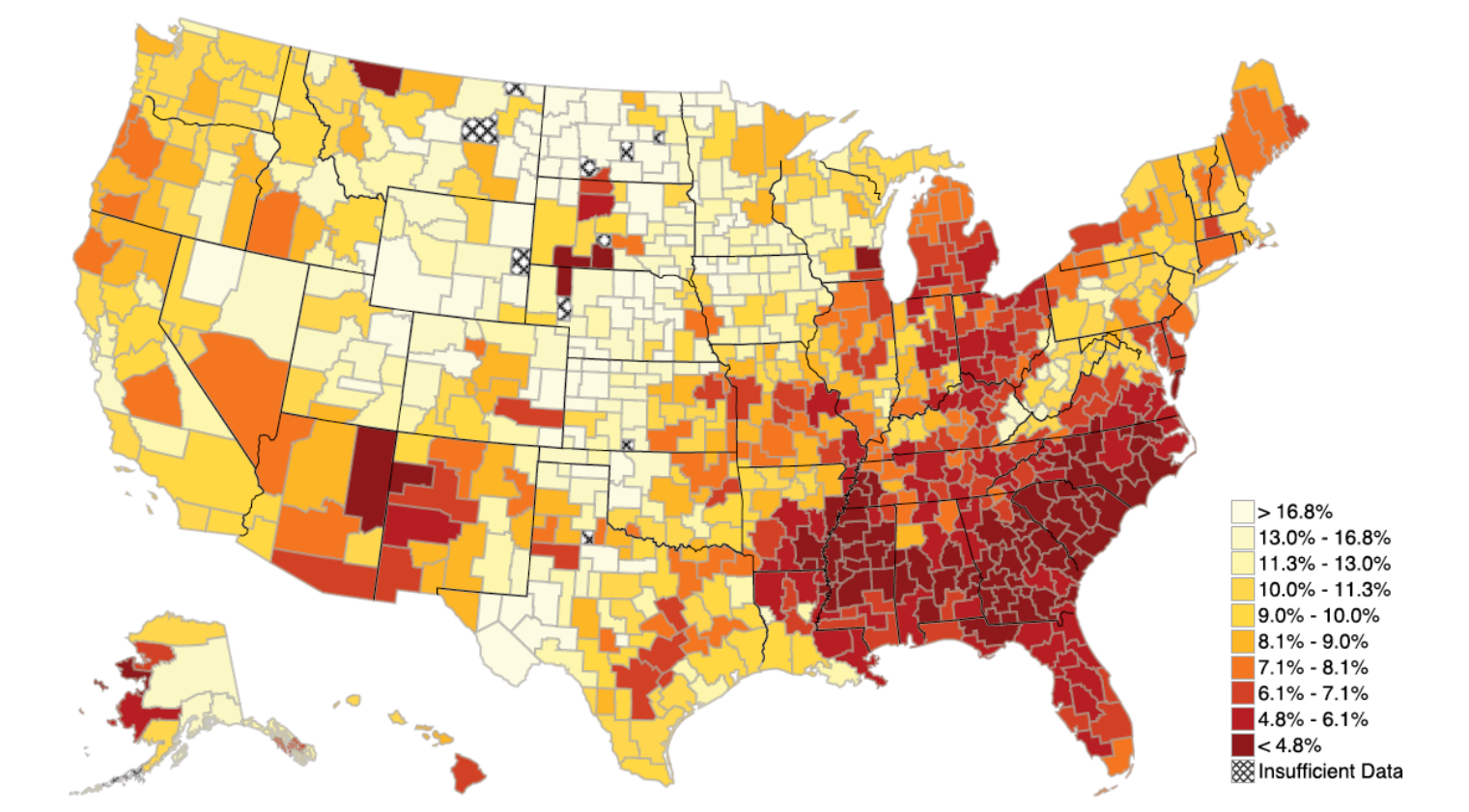


Figure 1: Economic mobility across US commuting zones

- Further applications include:
  - Evaluating what best improves test scores in public schools
  - Evaluating productivity differences between firms
  - Linking broadband access to productivity gains
- All made possible by tracking outcomes over long time periods
- Disadvantages:
  - Confidentiality and privacy issues
  - Limited compatibility between data sets

Private sector data

- Advantages:
  - Wide range of variables
  - Economically relevant
  - High speed
- Application: the Billion Prices Project

Private sector data: Billion Prices Project

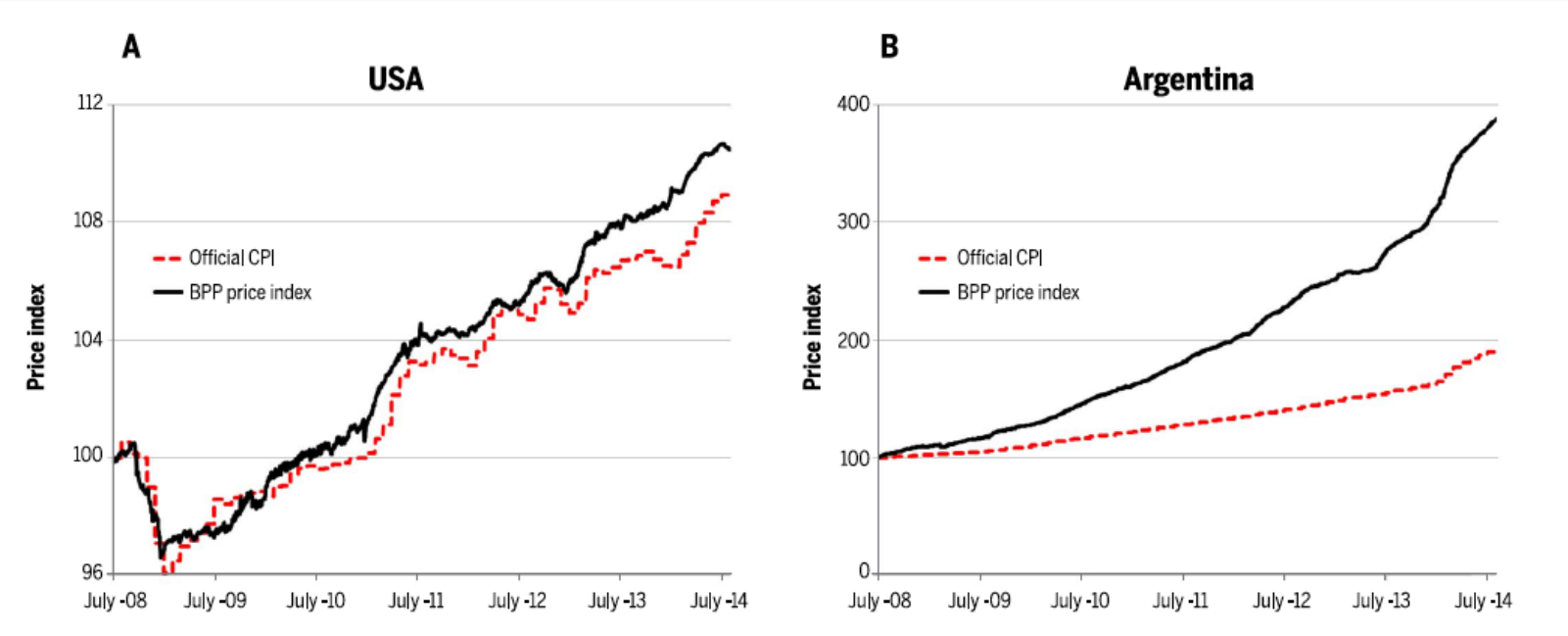


Figure 2: The BPP price index compared to the official price index

Private sector data: Billion Prices Project

- The Billion Prices Project constructs daily price indices based on real-time online data
- Good proxy for American Consumer Price Index (CPI)
- Potential alternative for unreliable official statistics?

Private sector data

- Other applications include:
  - Using newspaper text to proxy uncertainty of economic policies
  - Google Trends to forecast demand for goods, consumer confidence, and unemployment
  - eBay marketplace data to study consumer sensitivity to sales tax
- Key differences from administrative data:
  - Non-representative sample
  - Variables and data collection methods may vary over time

Private sector data

- Disadvantages:
  - Data access: even if data is shared, researchers have to keep it confidential → limited replication
  - Conflicts of interest

Statistical methods and role of theory

- Traditional econometric techniques and data mining for large-data differ fundamentally:
  1. causality vs prediction
  2. theory vs data driven modelling
  3. (focus on) statistical vs model uncertainty
- Idea: these approaches need not be in competition!

Statistical methods and role of theory

- Automated model/variable selection
- Better prediction $\rightarrow$ more tailored policies
- Economic theory is still important: large data sets need organising frameworks
- Econometric Theory still important: findings need careful interpretation/ causality still often indispensable

Einav & Levin (2014): conclusion

- Big data can help:
  - Better answer old questions
  - Pose interesting new questions
- There are several challenges:
  - Improving data capabilities
  - Coming up with new creative approaches for handling large data sets
- 'Big data is not a substitute for common sense, economic theory, or the need for careful research designs'

Prediction Policy Problems

Introduction

- Authors: Jon Kleinberg et al. (2015)
- Focus: clarify the distinction between causation and prediction
- Goal: show how both causation and prediction are policy-relevant

Idea: dealing with rain

- Consider two policymakers:
  1. A policymaker facing a drought must decide whether to invest in a rain dance
  2. A second policymaker must decide whether to take an umbrella to work
- Both policymakers benefit from knowing more about rain. However:
  1. One requires causality: do rain dances cause rain?
  2. The other requires prediction: is the chance of rain high enough to merit an umbrella?

Illustration: osteoarthritis

- Osteoarthritis is a painful chronic condition affecting the elderly
- Joint replacement surgery improves patient quality of life, however, there are costs:
  - Monetary costs: about $15,000
  - Non-monetary costs: pain, recovery time...
- Surgery only makes sense if the patient lives long enough to enjoy the benefits
- Determining whether to undergo surgery is a prediction problem.

Illustration: osteoarthritis

- Analysis on patients in the United States using Machine Learning (ML) identifies that many futile procedures occur in practice
- Avoiding such procedures could result in approx. $158 million being allocated elsewhere

- Similar analyses using machine learning tools include:
  - Detaining vs releasing arrestees based on predicted individual probability of committing crimes
  - Predicting which teachers add the most value in education systems
  - Predicting length of unemployment spells to tailor recommended job search strategies
  - etc.
- Prediction can also generate theoretical insights
  - e.g. investigating discrepancies between human and algorithmic decisions (behavioural economics)

Kleinberg et al. (2015): conclusion

- Prediction policy problems are important, common, and interesting
- Insights from machine learning should be adopted by policymakers for prediction problems

How to Tell When Simpler Theories Will Provide More Accurate Predictions

Introduction

- Authors: Malcolm Forster and Elliott Sober (1994)
- Focus: Solving prediction problems as curve fitting problems
- Goal: examine how the nature of the data should inform prediction problems
- Note: only the introductory part of the paper is covered

Idea

- Prediction can be viewed as a curve fitting problem

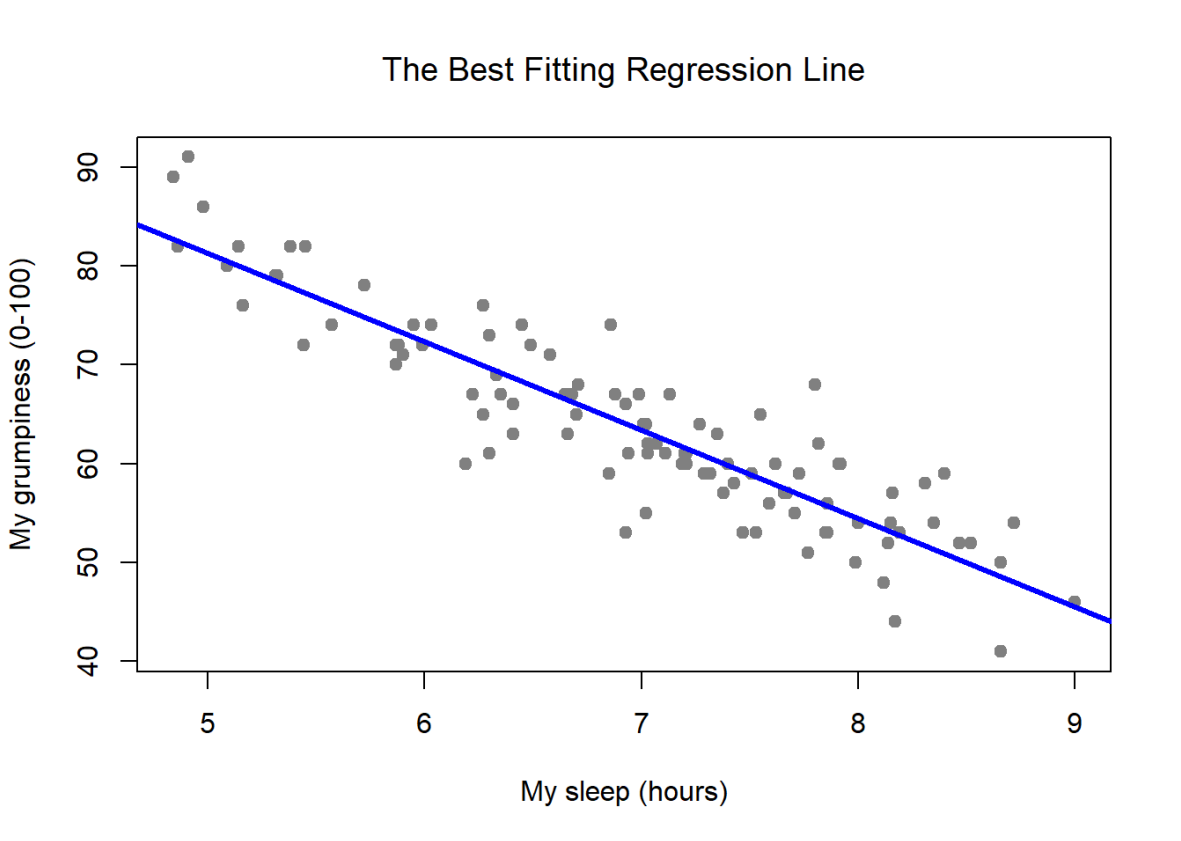


Figure 3: Example of curve fitting: hours of sleep and grumpiness

Idea

- Curve fitting consists of two steps:
  1. Determining the general shape of the curve (line, parabola, exponential...)
  2. Finding the parameters that make the curve best fit the data
- Step 1 defines how simple the model is
- Step 2 maximises the goodness-of-fit
- Trade-off between simplicity and goodness-of-fit

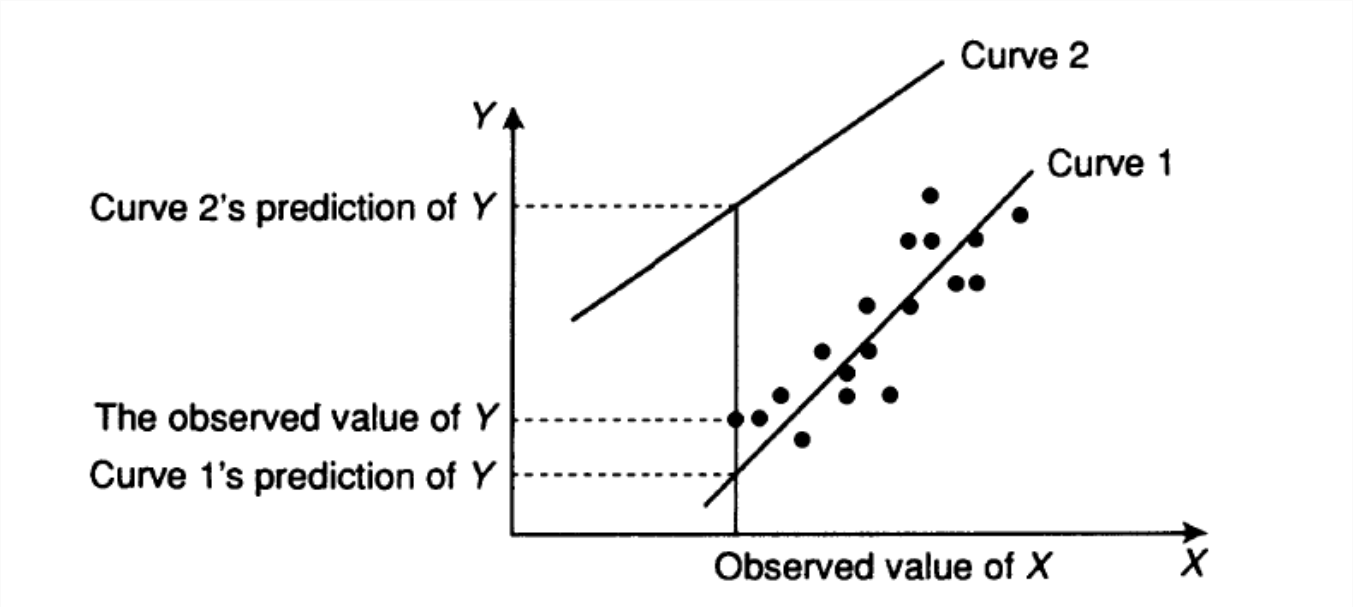


Figure 4: Equally simple curves with different levels of fit

- To minimise error, it is possible to fit a more complex curve that passes through all data points
- This makes no sense!
- But: relationships between variables can be non-linear

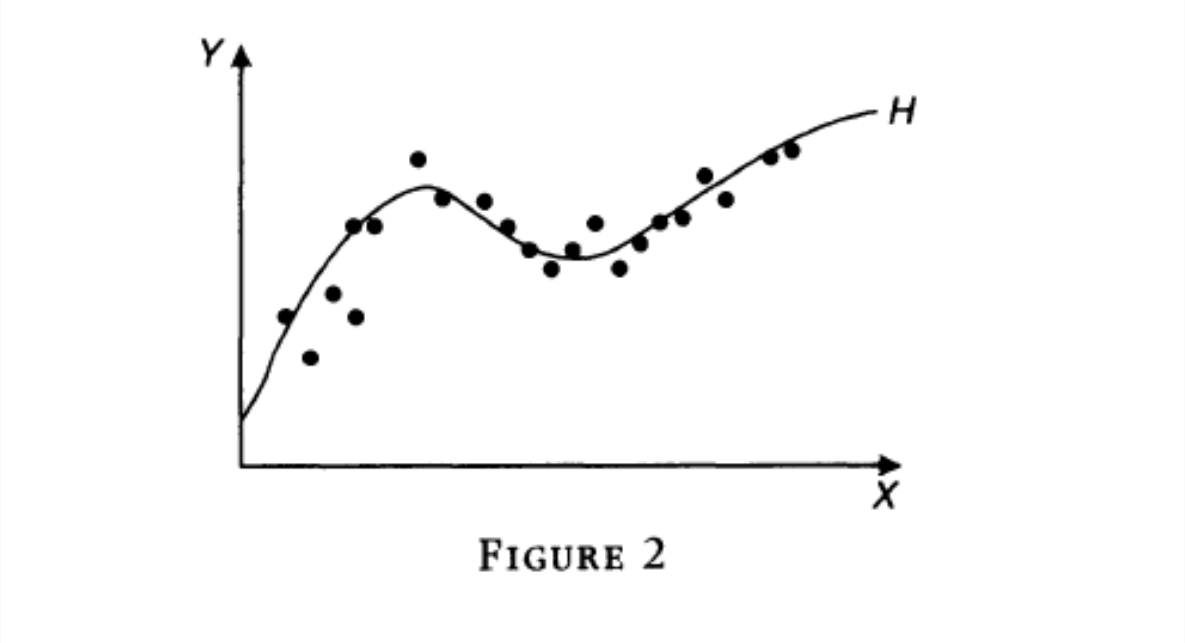


Figure 5: A non-linear relationship

Forster & Sober (1994): takeaway

- It is helpful to think of the 'true' curve as a signal and the error terms as noise
- Closeness to the data $\neq$ closeness to the truth
- Suggestions in solving the curve fitting problem include:
  - Consider simple curves first
  - Keep in mind that in-sample fit $\neq$ out-of-sample fit

Recap

You have learned

- ... what distinguishes Big Data and how Big Data challenges economists and econometricians in a positive way
- ... why (some) policy decisions can be treated as prediction problems and
- ... how we can illustrate the trade-off between model simplicity and model complexity in the case of the curve fitting prolem